

**Institute of Mathematical Statistics**  
**LECTURE NOTES—MONOGRAPH SERIES**

---

# **$L_1$ -Statistical Procedures and Related Topics**

**Yadolah Dodge, Editor**

---





**Institute of Mathematical Statistics**

**LECTURE NOTES–MONOGRAPH SERIES**

**Volume 31**

# **$L_1$ -Statistical Procedures and Related Topics**

**Yadolah Dodge, Editor**

**Institute of Mathematical Statistics  
Hayward, California**

Institute of Mathematical Statistics

*Lecture Notes-Monograph Series*

Editorial Board

Andrew A. Barbour, Joseph Newton, and David Ruppert (Editor)

The production of the *IMS Lecture Notes-Monograph Series* is managed by the IMS Business Office: Miriam Gasko Donoho, IMS Treasurer, and James H. Sanders, IMS Business Manager.

Library of Congress Catalog Card Number: 97-073621

International Standard Book Number 0-940600-43-9

Copyright © 1997 Institute of Mathematical Statistics

All rights reserved

Printed in the United States of America

Dedicated to the memory of  
*Roger Joseph Boscovich*  
Father of  $L_1$  methods



Roger Joseph Boscovich (1711 - 1787)



# Preface

This volume contains a selection of papers presented to the Third International Conference on the  $L_1$ -Norm and Related Methods, held in Neuchâtel, Switzerland, from August 11-15, 1997, as a Satellite Meeting to the 51st ISI Session in Istanbul. The conference included invited talks, contributed papers and a tutorial. A Summer School in Regression and Time Series Analysis for young graduate students and research workers ran in parallel.

The success of the 1987 and 1992 conferences on the Statistical Data Analysis based on the  $L_1$ -Norm and related methods made it evident that there is a need for regular conferences on the topic. For this reason we launched the third and happily brought together many new faces, especially those of younger statisticians.

This volume includes 38 invited papers listed under nine headings.

The Prologue contains two papers. One on measuring the performance of boundary-estimation methods by Peter Hall and Marc Raimondo and another by Roger Koenker on a new computational procedure for  $L_1$ . These two papers make the opening and closing lectures of the conference. Peter Hall and Marc Raimondo consider the problem of linear approximation to a curved boundary using a gridded data which is closely related to both curve estimation in statistics and rational approximation in number theory. They show that measures of performances based on the  $L_1$  norm are more appropriate for the problem than those found in  $L_p$  norms for  $p > 1$ . Roger Koenker's breakthrough in computation of  $L_1$  is rather different from the simplex method since it does not iterate around the exterior of the constraint set. When there are many observations, the simplex algorithm becomes too slow in computation since it has to pass through too many vertices to achieve the optimal solution. His algorithm starts in the interior of the constraint set and does penalized Newton steps with the log-barrier formulation designed to keep the algorithm in the interior.

Part one contains seven papers on estimation, testing and characterization. A new regression rank statistic for testing general hypothesis in a class of non-parametric linear model is introduced by Cornelius Gutenbrunner. In the same article, Gutenbrunner develops the asymptotic representation



of his regression rank statistics. Marc Hallin and Ivan Mizera study the unimodality and asymptotics of M-estimates. Marie Hušková, after describing the likelihood principle for the normally distributed errors, provides the reader with four  $L_1$  type tests for the change point problem. The likelihood ratio, a Wald type, a score type and Bayesian type test statistics. She then studies the limit behavior of these tests statistics under the null hypothesis. Stephan Morgenthaler investigates the behaviour of residuals from  $L_1$  fit in linear models of designed experiments. Stephan argues that the residuals obtained by  $L_1$  fitting in classification models exhibit some weaknesses. Jana Jureckova, Klebanov and Silvelyn 20 (her signature) contribute two interesting papers on the inadmissibility of robust estimators with respect to the  $L_1$ -norm and the  $L_1$  estimation in nonlinear regression and in nonlinear error-in-variables models. For testing the hypothesis  $H_0 : L_\beta = L$  versus  $H_1 : L_\beta \neq L$  in linear models, Christine Müller derives a Wald type test statistic based on the  $L_1$ -estimator with maximum relative power under the side condition that the maximum bias is bounded by some bias bound. A one-way classification model is given to clarify her  $L_1$  test statistics.

Part two is on computational procedures, algorithms and computer packages. José Agullo proposes two new exact algorithms for the least median of squares and Mia Hubert and Peter Rousseeuw put forward new developments in their computer package PROGRESS. Bill Farebrother outlines the early history of traditional estimation of observational equations from the third century to Roger Boscovich passing through the Tobias Mayer method. Steve Portnoy explains his new findings on the computation of  $L_1$  methods and shows that his method of computing is faster than computing least squares for large data sets with  $n = 10^4$  observations. Steve together with Roger achieved their goal by replacing the simplex approach with interior point methods based on a stochastic preprocessing step that begins with a much smaller random subset of the data. I am sure that with this breakthrough algorithm the area of  $L_1$  statistical data analysis takes a new turn in the future. Chris Adcock, Meade, Chepoi, Cogneau, Bernard Fichet and Fitzenberger add their new findings to this part and provide directions for future developments.

Statistical graphics are grouped together in Part three. Efstathia Bura, Simon Sheather and Joseph McKean present new ideas with the application of inverse regression for dimension reduction and Francesca Chiaramonte provides the foundation and philosophical reasoning behind a reduction paradigm. Elias Moreno proposes a Bayesian method of model selection. Wetzel describes key features of interactive method which construct CERES plots, a new class of plots that includes the partial residual plots for iden-

tification of curvature in regression models.

Time series analysis and financial statistics are the subject of Part four of this volume. Five articles by Gib Bassett, Hurst and Platen, Keith Knight, Hans Nyquist and Terui and Kariya discuss  $L_1$  and  $L_2$  methods for the analysis of time and money!

Bruce Brown, Tom Hettmansperger, Möttönen and Oja present the rank plot in the affine invariant case in Part five and Fernholz, Stute and Una-Alvarez, Gonzalez-Manteiga and Cadarso-Suarez contribute to this non-parametric part with new ideas in target estimation, model check and density estimation.

Part six contains three contributions on multivariate analysis. Biman Chakraborty and Probal Chaudhuri study an extension of rank regression techniques to multivariate models. Philip Milasevic and Deb Nolan present an original paper on mode and concentration estimation in multidimensions. Nolan dedicated her joint work to the memory of her co-author Milasevic. By means of empirical process theory, Deb and Milasevic provide the rate of convergence and limiting distributions of their estimators. Fraiman, Regina Liu and Jean Meloche study the estimation of multivariate density by probing depth. They consider a class of multivariate densities within which a density function  $f$  can be expressed as  $f = g \circ D$  in some notion of data depth  $D$  and some real function  $g$ .

In Part seven, three papers by Anil Chaturvedi and Doug Carroll, Laurence Hubert, Phipps Arabie and Jacqueline Meulman and Boris Mirkin discuss different approaches with  $L_p$ -norm to classification problems.

Fourteen of the fifty two invited papers appear at the end of volume in abstract form as they are published elsewhere.

If a volume or a manuscript has a prologue it should, in principle, have an epilogue. The volume is about  $L_1$  methods in all area of statistics. As Roger Koenker stated in his paper in the prologue section: "Despite the best efforts of such distinguished advocates as Laplace (1789), Edgeworth (1888), and Kolmogorov (1931), methods of estimation based on minimizing sums of absolute errors have languished in the shade of the edifice that Gauss built on the foundation of least squares. Why?". The volume would not have been complete without some prognosis for the future of robust methodologies. This is why I took the liberty to call Peter Huber and asked him his permission to include his paper *Robustness: Where are we now ?* as an epilogue. I had remembered the story:

Devil: *Master, Master, we are in serious trouble.  
On the planet Earth, Man has discovered the truth.*

Master: *Don't worry, truth will soon become dogma.*

To assure high standards, all papers were subjected to an intensive refereeing process involving at least one and often two referees. The papers once chosen were again subject to the Series Editor's selection. Based upon space and time not all papers submitted were accepted for publication. I am grateful to the following for their conscientious refereeing:

| Name                      | Affiliation                                       |
|---------------------------|---------------------------------------------------|
| José Agullo               | Universidad de Alicante (Spain)                   |
| Gérard Antille            | Université de Genève (CH)                         |
| Gib Basset                | University of Illinois (USA)                      |
| Efstathia Bura            | George Washington University (USA)                |
| J. Douglas Carroll        | Rutgers University (USA)                          |
| Victor Chepoi             | Lab. De Biomathématique, Marseille (France)       |
| Francesca Chiaromonte     | Int. Inst. for Applied Systems Analysis (Austria) |
| Dennis Cook               | University of Minnesota (USA)                     |
| David Draper              | University of Bath (UK)                           |
| Jin-Chuan Duan            | Hong Kong University of Science and Technology    |
| Michel Dubois             | Université de Neuchâtel (CH)                      |
| Bernard Fichet            | Lab. De Biomathématique, Marseille (France)       |
| Bernd Fitzenberger        | Universität Konstanz (Germany)                    |
| Riccardo Gatto            | Université de Neuchâtel, Neuchâtel (CH)           |
| Wences. Gonzalez-Monteiga | University Santiago de compostela (Spain)         |
| Marc Hallin               | Université Libre de Bruxelles (Belgium)           |
| Willem Heiser             | Leiden University (NL)                            |
| Stéphane Héritier         | Hôpital universitaire de Genève (CH)              |
| Tom Hettmansperger        | Penn State University, PA (USA)                   |
| Mia Hubert                | University of Antwerp (Belgium)                   |
| Simon R. Hurst            | Australian National University (Australia)        |
| Marie Huskova             | Charles University, Prague (CZ)                   |
| Jana Jureckova            | Charles University, Prague (CZ)                   |
| Keith Knight              | University of Toronto (Can)                       |
| Roger Koenker             | University of Illinois (USA)                      |
| Regina Liu                | Rutgers University (USA)                          |
| F.R. Mc Morris            | University of Louisville (USA)                    |
| Joe McKean                | Western Michigan University (USA)                 |
| Jean Meloche              | University of British Columbia (Can)              |
| Jacqueline Meulman        | University of Leiden (NL)                         |
| Boris Mirkin              | German Cancer Research Center (Germany)           |
| Elias Moreno              | University of Granada (Spain)                     |
| Stephen Morgenthaler      | Swiss Federal Institute of Technology (CH)        |
| Christine Müller          | Freie Universität Berlin (Germany)                |
| Hans Nyquist              | University of Umea (Sweden)                       |
| Wolfgang Polasek          | Inst. of Statist. and Econometrics, Basel (CH)    |

| Name              | Affiliation                                    |
|-------------------|------------------------------------------------|
| Stephen Portnoy   | University of Illinois (USA)                   |
| Peter Rousseeuw   | University of Antwerp (Belgium)                |
| Allan Seheult     | University of Durham (UK)                      |
| Simon Sheather    | University New South Wales (Australia)         |
| Arnold Stromberg  | University of Kentucky (USA)                   |
| Maurizio Vichi    | Università "G.D'Annunzio" di Chieti (Italy)    |
| Joe Whittaker     | Lancaster University (UK)                      |
| Lincheng Zhao     | University of Science & Technology, Hefei (CN) |
| Silvelyne Zwanzig | Universität Hamburg (Germany)                  |

Organising and hosting a conference even at this modest level, requires lot of courage, patience, and above all a responsible secretary. I was very fortunate to have a very dedicated one. I wish to express my gratitude to Melanie Miserez for handling with extreme care all the correspondence. I also owe a great debt of gratitude to Gérard Geiser for his outstanding TEX skill in the production of the volume and to Valentin Rousson who generously worked along with Gérard to make the production faster.

I wish to acknowledge the generous support of the Swiss National Science Foundation (Grant No 2101-49'913.96), the Swiss Academy of Humanities and Social Sciences and the University of Neuchâtel. Without the financial aid of these agencies the conference could not have been held.

I am grateful to all those who participated the conference, and to the organizers of the invited sessions Dennis Cook, Willem Heiser, Jana Jureckova, Regina Liu, Joe McKean, Stephan Morgenthaler, Hans Nyquist, Wolfgang Polasek, Steve Portnoy, Peter Rousseeuw, Gabreila Stangenhäus, Takeaki Kaiya, Maurizio Vichi, Jinde Wang, and Joe Whittaker, and certainly, to the invited persons whose contribution made this volume possible.

Finally, on the other side of the Atlantic, David Ruppert, James Sanders and Patti Shankland, did not stop their encouragement. Having the volume ready in time represents a considerable investment of their efforts and I do thank them very sincerely.

It takes courage, good staff and a great number of colleagues to organize a conference. I certainly would not recommend this burden to anyone but the insane.

University of Neuchâtel  
Switzerland  
August 1997

Yadolah Dodge  
Editor



# Table of contents

|                                                                                                                       |      |
|-----------------------------------------------------------------------------------------------------------------------|------|
| <b>PREFACE</b> .....                                                                                                  | iii  |
| <b>INTRODUCTION</b> .....                                                                                             | xiii |
| <b>PROLOGUE</b>                                                                                                       |      |
| Measuring the performance of boundary-estimation methods<br>P. Hall and Marc Raimondo .....                           | 1    |
| $L_1$ computation: An interior monologue<br>R. Koenker .....                                                          | 15   |
| <b>I. ESTIMATION, TESTING AND CHARACTERIZATION</b>                                                                    |      |
| Regression rank statistics<br>C. Gutenbrunner.....                                                                    | 33   |
| Unimodality and the asymptotics of M-estimators<br>M. Hallin and I. Mizera .....                                      | 47   |
| $L_1$ -test procedures for detection of change<br>Marie Huskova .....                                                 | 57   |
| Inadmissibility of robust estimators with respect to $L_1$ norm<br>J. Jurecková and L.B. Klebanov .....               | 71   |
| Properties of $L_1$ -residuals<br>S. Morgenthaler .....                                                               | 79   |
| $L_1$ -test in linear models: Tests with maximum relative power<br>C. Müller .....                                    | 91   |
| On $L_1$ -norm estimators in nonlinear regression and in nonlinear<br>error-in-variables models<br>S. Zwanzig.....    | 101  |
| <b>II. COMPUTATIONAL PROCEDURES AND ALGORITHMS</b>                                                                    |      |
| A comparison of two LP Solvers and a new IRLS Algorithm for<br>$L_1$ estimation<br>C. Adcock and N. Meade .....       | 119  |
| Exact algorithmes to compute the least median of squares estimate<br>in multiple linear regression<br>J. Agullo ..... | 133  |

|                                                                                 |     |
|---------------------------------------------------------------------------------|-----|
| Polynomial algorithms for isotonic regression                                   |     |
| V. Chepoi, D. Cogneau and B. Fichet .....                                       | 147 |
| Notes on the early history of elemental set methods                             |     |
| R.W. Farebrother .....                                                          | 161 |
| Computational aspects of censored quantile regression                           |     |
| B. Fitzenberger .....                                                           | 171 |
| On computation of regression quantiles: Making the Laplacian<br>Tortoise faster |     |
| S. Portnoy .....                                                                | 187 |
| Recent developments in PROGRESS                                                 |     |
| P.J. Rousseeuw and M. Hubert .....                                              | 201 |

### **III. STATISTICAL GRAPHICS: FOUNDATIONS AND METHODOLOGY**

|                                                                                            |     |
|--------------------------------------------------------------------------------------------|-----|
| Dimension reduction via parametric inverse regression                                      |     |
| E. Bura .....                                                                              | 215 |
| A reduction paradigm for multivariate laws                                                 |     |
| F. Chiaromonte .....                                                                       | 229 |
| Exploring data sets using partial residual plots based on robust fits                      |     |
| J. W. McKean and S. J. Sheather .....                                                      | 241 |
| Bayes factors for intrinsic and fractional priors in nested models.<br>Bayesian robustness |     |
| E. Moreno .....                                                                            | 257 |
| A comparison of procedures based on inverse regression                                     |     |
| S. J. Sheather and J. W. McKean .....                                                      | 271 |
| Interactive graphical modeling methods for regression                                      |     |
| N. Wetzel .....                                                                            | 279 |

### **IV. TIME SERIES ANALYSIS AND FINANCIAL STATISTICS**

|                                                                                     |     |
|-------------------------------------------------------------------------------------|-----|
| Nonparametric bounds for the probability of future prices based<br>on option values |     |
| G. W. Bassett Jr. ....                                                              | 287 |
| The marginal distributions of returns and volatility                                |     |
| S.R. Hurst and E. Platen .....                                                      | 301 |

|                                                                                              |     |
|----------------------------------------------------------------------------------------------|-----|
| Some limit theory for $L_1$ -estimators in autoregressive models<br>under general conditions |     |
| K. Knight .....                                                                              | 315 |
| A Lagrange multiplier approach to testing for serially dependent<br>error terms              |     |
| H. Nyquist .....                                                                             | 329 |
| Gaussianity and nonlinearity of foreign exchange rates                                       |     |
| N. Terui and T. Kariya .....                                                                 | 337 |

## V. NONPARAMETRIC INFERENCE

|                                                                                                      |     |
|------------------------------------------------------------------------------------------------------|-----|
| Rank plots in the affine invariant case                                                              |     |
| B. M. Brown, T.P. Hettmansperger, and J. Möttönen and H. Oja ..                                      | 351 |
| Target estimation and implications to robustness                                                     |     |
| L. Fernholz .....                                                                                    | 363 |
| Model checks in statistics: An innovation process approach                                           |     |
| W. Stute .....                                                                                       | 373 |
| Bootstrap selection of the smoothing parameter in density estimation<br>under the Koziol-Green model |     |
| J. de Una-Alvarez, W. Gonzalez-Manteiga and C. Cadarso-Suarez .                                      | 385 |

## VI. MULTIVARIATE ANALYSIS

|                                                                                    |     |
|------------------------------------------------------------------------------------|-----|
| On multivariate rank regression                                                    |     |
| B. Chakraborty and P. Chaudhuri .....                                              | 399 |
| Multivariate density estimation by probing depth                                   |     |
| R. Fraiman, R.Y. Liu and Jean Meloche .....                                        | 415 |
| Modes, caps and concentration: A geometric approach to estimation<br>on the sphere |     |
| P. Milasevic and D. Nolan .....                                                    | 431 |

## VII. CLASSIFICATION

|                                                                                              |     |
|----------------------------------------------------------------------------------------------|-----|
| An $L_1$ -norm procedure for fitting ADCLUS and INDCLUS                                      |     |
| A. Chaturvedi and J. D. Carroll .....                                                        | 443 |
| Hierarchical clustering and the construction of (optimal) ultrametrics<br>using $L_p$ -norms |     |
| L. Hubert, P. Arabie and J. Meulman .....                                                    | 457 |



$L_1$  and  $L_2$  approximation clustering for mixed data: Scatter  
decompositions and algorithms  
    B. Mirkin.....473

**EPILOGUE**

Robustness: Where are we now ?  
    P. J. Huber.....487

**ABSTRACTS**.....xix

**CONTRIBUTORS**.....xxv

# Introduction

**Yadolah Dodge**

*University of Neuchâtel, Switzerland*

While the method of least squares (and its generalizations) have served statisticians well for a good many years (mainly because of mathematical convenience and ease of computation), and enjoys certain well known properties within strictly Gaussian parametric models, it is recognized that outliers, which arise from heavy-tailed distributions, have an unusually large influence on the estimates obtained by these methods. Indeed, one single outlier can have an arbitrary large effect on the estimate. Outlier diagnostics have been developed to detect observations with a large influence on the least squares estimation. For excellent books related to such diagnostics the reader is referred to Cook and Weisberg (1982, 1994) and Chatterjee and Hadi (1988).

Parallel to diagnostic techniques, robust methods with varying degrees of robustness and computational complexity have been developed to modify the LS method so that the outliers have less influence on the final estimates. Among others are the bounded influence estimators, the repeated median, the least median of squares and the regression quantile methods.

In 1964, Huber published what is now considered to be a classic paper on robust estimation of location parameter and subsequently extended to that linear model. The development of selected robustness concepts since their inception in the 1960's and their current status, is given by Huber (1995).

One of the simplest robust alternatives to LS is the least absolute value method. This method, which is the subject of this volume, is a widely recognized superior method especially well-suited to longer-tailed error distributions, such as the Laplace distribution.

Depending on the field of application, the least absolute value method has been studied in several contexts under a variety of names such as minimum, or least sums of absolute errors, deviations or values; and here we

refer to it as the  $L_1$ - norm method (for minimizing the  $L_1$ - norm of the vector of deviations). The  $L_1$  method estimates the unknown parameters in a stochastic model so as to minimize the sum of the absolute deviations of a given set of observations from the values predicted by the model.

Historically,  $L_1$  estimation is the oldest of all robust methods. The method of least absolute deviations was introduced almost 50 years before the method of least squares, in 1757 by Roger Joseph Boscovich (1711-1787). He devised the method as a way to reconcile inconsistent measurements for estimating the shape of the earth. After Pierre Simon, Laplace adopted the method 30 years later, it saw occasional use but was soon overshadowed by the method of least squares. The popularity of least squares was at least partly due to the relative simplicity of its computations and to the supporting theory that was developed by Gauss and Laplace. Laplace, in his second memoir on the Figure of the Earth in 1789, adopted Boscovich's two criteria for a line of best fit, and gave an algebraic formulation and derivation of Boscovich's algorithm.

After nearly seventy years following the publication of Laplace's second supplement to the *Théorie Analytique des Probabilités* (1818), Edgeworth (1887) presented a method for linear regression using  $L_1$  method. But since the publication of Edgeworth's work, few attempts have been made to convince the statisticians and particularly the applied users to employ this method (see Turner, 1887; Rhodes, 1930; Singleton, 1940; Karst, 1958). Reasons for such a long silence may be summarized as follows :

(1) Computational difficulties in producing the numeric values of the  $L_1$  estimates in regression. (Lack of closed form formulae similar to that of least squares).

(2) Lack of an asymptotic theory for  $L_1$  estimation in the regression model, and more generally the nonexistence of accompanying statistical inference procedures.

(3) Insufficient evidence to show the superiority of the small sample properties of  $L_1$  estimation compared to the LS estimators when sampling from long tailed distributions.

Following the work of Charnes, Cooper and Ferguson (1955) a renewed interest in using  $L_1$  estimation for regression problem was created. They showed the equivalence between the  $L_1$  problem and a linear programming problem. Wagner (1959) suggested that the  $L_1$  problem in a linear regression of the form  $y_i = \theta_0 + \theta_1 x_{1i} + \dots + \theta_p x_{pi} + \varepsilon_i$  or in matrix form  $Y = X\theta + \varepsilon$  can be solved by solving the dual of the  $L_1$  problem. He also observed that the dual problem can be reduced to a problem with a smaller basis but where the dual variables have upper-bound restrictions. Wagner's formulation of the problem is to restate the problem of minimizing  $\sum |\varepsilon_i|$  with

respect to  $\theta$  where  $\varepsilon_i$  is the deviation between the observed and predicted values of the  $i^{th}$  observation  $Y_i$ , as: minimize  $\sum |\varepsilon_i|$ , subject to  $X\theta + \varepsilon = Y$ , where  $\theta, \varepsilon$  unrestricted in sign.

Noting the fact that  $|\varepsilon_i| = \varepsilon_{1i} + \varepsilon_{2i}$  where  $\varepsilon_1 = \varepsilon$  if  $\varepsilon > 0$ , 0 otherwise and  $\varepsilon_2 = -\varepsilon$  if  $\varepsilon < 0$ , 0 otherwise, that both are nonnegative and  $\varepsilon_i = \varepsilon_{1i} - \varepsilon_{2i}$ , we can reformulate the problem as a linear programming problem: minimize  $\sum \varepsilon_{1i} + \sum \varepsilon_{2i}$ , subject to  $X\theta + \varepsilon_1 - \varepsilon_2 = Y$ , where  $\theta$  unrestricted in sign,  $\varepsilon_1, \varepsilon_2 \geq 0$ .

From the computational point of view, the  $L_1$  method is now extremely simple and it requires only a routine to fit the  $L_1$  regression. There are several computer programs available for calculation of  $L_1$  estimates. See for example Sadoski (1974) and Farebrother (1988). For the case of multiple regression we can use, for example, the modified simplex algorithm of Barrodale and Roberts (1973) that exists in the IMSL library under the name RLLAV. The  $L_1$  estimation problem with additional linear restrictions (restricted  $L_1$  problem) is considered along the same lines in Barrodale and Roberts (1974). Arthanari and Dodge (1993) devoted a complete chapter on computational aspects of the  $L_1$  estimation.  $L_1$  regression estimates are also obtainable from the function `l1fit` in the computer language S-Plus and from the robust regression package ROBSTATS (Marazzi, 1993). Detection of outlying points in both dependent and independent variables in regression model are explained in Dodge (1997).

The major difficulty for applied researchers in using  $L_1$  estimation for many years was the lack of accompanying statistical inference procedures. Such procedures would include methods for testing general linear hypothesis, obtaining confidence intervals, analysis of variance tables and for performing multiple comparison procedures.

Bassett and Koenker (1978) developed the asymptotic theory for  $L_1$  estimators in the regression model. Their finding is considered to be a breakthrough for the problem. Their main result is that the sampling distribution of  $L_1$  estimators will be asymptotically normal with a specified mean and variance. Under very general assumptions they confirmed that the  $L_1$  estimator  $\hat{\theta}$  has a normal distribution with mean  $\theta$  and covariance matrix  $\tau^2(X'X)^{-1}$  where  $\tau^2/n$  is the asymptotic variance of the sample median from random samples of size  $n$  taken from the error distribution with a continuous and positive derivative at the median. This result is remarkably similar to that for LS. Therefore the  $L_1$  confidence intervals for an estimable function  $\lambda'\hat{\theta}$  is

$$\lambda'\hat{\theta} \pm z_{\alpha/2} \cdot \hat{\tau} \sqrt{\lambda'(X'X)^{-}\lambda}$$

where  $(X'X)^{-}$  is the q-inverse of  $X'X$  and  $\hat{\tau}$  is an estimate of  $\tau$  given in

McKean and Schrader (1987).

Koenker and Bassett (1982) investigate the asymptotic distribution of three alternative  $L_1$  test statistics of a linear hypothesis in the standard linear model. They showed that the three test statistics, which correspond to the Wald, likelihood ratio and Lagrange multiplier tests, under mild regularity conditions on the design and error distribution, have the same limiting chi-square behavior. For a complete treatment of  $L_1$  regression the reader is referred to Chapter 4 of Birkes and Dodge (1993).

With the availability of many computationally efficient algorithms and developed inference procedures for testing general linear hypotheses, for obtaining confidence intervals, selection of variables, analysis of variance tables and multiple comparison, it is hoped that  $L_1$  estimation methods will be employed more often by researchers in applied sciences than before.

Certainly, there are many other areas of statistical data analysis based on the  $L_1$ -norm (such as density estimation, time series analysis, multivariate analysis and classification methods) that could have been discussed here. But, unfortunately, limitation of space and time have caused many interesting and important lines of research to be treated lightly or not at all. While it is now evident that no single robust procedure is best by any criteria, it may be appropriate (or at least reasonable) to use adaptive convex combinations of  $L_1$  with other methods rather than a single criterion to estimate the unknown parameters. However, for the error distributions for which the median is superior to the mean as an estimator of location,  $L_1$  estimation is certainly preferred to least squares and strongly recommended for use in these cases.

Bloomfield and Steiger (1983), Devroye and Györfi (1985), and Gonin and Money (1989) are the only books entirely devoted to  $L_1$  topic. The authors of these three texts had the courage to pull together a rich and diverse literature in this field. I hope that the proceedings contained in this volume and its predecessors, Dodge (1987, 1992), will encourage someone to write a fourth.

## References

- [1] Arthanari, T.S. and Dodge, Y. (1993). *Mathematical Programming in Statistics*. New York: Wiley, Classical Edition.
- [2] Barrodale, I. and Roberts, F.D.K. (1973). An improved algorithm for discrete  $L_1$  linear approximation. *SIAM J. Numer. Anal.* **10**, 839-848.
- [3] Barrodale, I. and Roberts, F.D.K. (1974). Algorithm 478: Solution of an overdetermined system of equations in the  $L_1$  norm. *Commun.*

- Assoc. Comput. Mach.* **17**, 319.
- [4] Bassett, G.W. and Koenker, R. (1978). Asymptotic theory of least absolute error. *J. Am. Statist. Assoc.* **73**, 618-22.
  - [5] Birkes, D. and Dodge, Y (1993). *Alternative Methods of Regression*. New York: Wiley.
  - [6] Bloomfield, P. and Steiger, W.L. (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms*. Boston: Birkäuser.
  - [7] Boscovich, R.J. (1757). De litteraria expeditione per pontificiam synopsis amplioris operis, ac habentur plura eius ex exempla ria etiam sensorum impressa. Bononiensi Scientiarum et Artium Instituto Atque Academia Commentarii 4, 353-96.
  - [8] Charnes, A., Cooper, W.W. and Fergusson, R.O. (1955). Optimal estimation of executive compensation by linear programming. *Manage. Sci.* **1**, 138.
  - [9] Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression*. New York: Wiley.
  - [10] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
  - [11] Cook, R.D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.
  - [12] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation, the  $L_1$  View*. New York: Wiley.
  - [13] Dielman, T.E. and Rose, L.E. (1995). A bootstrap approach to hypothesis testing in least absolute value regression. *Comp. Statist. Data Anal.* **20**, 119-130.
  - [14] Dielman, T.E. and Rose, L.E. (1996). A note on the hypothesis testing in LAV multiple regression: A small sample comparison. *Comp. Statist. Data Anal.* **21**, 463-470.
  - [15] Dodge, Y. (1987). (Ed). *Statistical Data Analysis Based on the  $L_1$ -Norm and Related Methods*. Amsterdam: North-Holland.
  - [16] Dodge, Y. (1992).  *$L_1$ -Statistical Data Analysis and Related Methods*. Amsterdam: North Holland.
  - [17] Dodge, Y. (1997). LAD regression for detecting outliers in response and explanatory variables. *J. Multiv. Anal.* **61**, 144-158.
  - [18] Edgeworth, F.Y. (1887). On observations relating to several quantities. *Phil. Mag.* (5th Series) **24**, 222.
  - [19] Ellis, S.P. and Morgenthaler, S. (1992). Leverage and breakdown in  $L_1$  regression. *J. Am. Statist. Assoc.* **87**, 143-148.
  - [20] Farebrother, R.W. (1988). A simple recursive procedure for the  $L_1$  norm fitting of a straight line. *Appl. Statist.* **37**, 457-489.
  - [21] Gonin, R. and Money, A.H. (1989). *Nonlinear  $L_p$ -Norm Estimation*.

New York: Marcel Dekker.

- [22] Huber, P.J. (1987). The place of the  $L_1$ -norm in robust estimation. In *Statistical Data Analysis - Based on  $L_1$ -Norm and Related Methods*, Ed. Y. Dodge, pp. 23-33. Amsterdam: North Holland.
- [23] Huber, P.J. (1995). Robustness: Where are we now ? *Student* **1**, 75-86.
- [24] Karst, O.J. (1958). Linear curve fitting using least deviations. *J. Am. Statist. Assoc.* **53**, 118-132.
- [25] Koenker, R. and Bassett, G.W. (1982). Tests of hypotheses and  $L_1$  estimation. *Econometrica* **50**, 1577-83.
- [26] Koenker, R. (1987). A comparison of asymptotic testing methods for  $L_1$ -regression. In this volume.
- [27] Laplace, P.S. (1789). Sur les degrés mesurés des méridiens, et sur les longueurs observées sur pendule. Histoire de l'Académie royale des inscriptions et belles lettres, avec les Memoires de littérature tirez des registres de cette académie, Année 1789. Paris. "Mémoires", 18-43.
- [28] Laplace, P.S. (1793). Sur quelques points du système du monde. Mémoires de l'Académie Royale des Sciences de Paris, 1-87. Reprinted in Oeuvres Complètes de Laplace, Vol.11, pp. 477-558. Paris: Gauthier-Villars, 1895.
- [29] Laplace, P.S. (1818). Deuxième supplément to Laplace (1812).
- [30] Legendre, A.M. (1805). Nouvelles méthodes pour la détermination des orbites des comètes. Paris: Courcier. Appendice sur la méthode des moindres carrés, 72-80.
- [31] Marazzi, A. (1993). *Algorithms, Routines, and S Functions for Robust Statistics*. Wadsworth and Brooks/Cole, California.
- [32] McKean, J.W. and Schrader, R.M. (1987). Least absolute errors analysis of variance. In *Statistical Data Analysis Based on the  $L_1$ -Norm and Related Methods*, Ed. Y. Dodge, pp 297-321. Amsterdam: North Holland.
- [33] Rhodes, E.C. (1930). Reducing observations by the methode of minimum deviations. *Phil. Mag.* (7th Series) **9**, 974.
- [34] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
- [35] Sadowski, A.N. (1974). Algorithm AS74.  $L_1$ -norm fit of a straight line. *Appl. Statist.* **23**, 244-248.
- [36] Singleton, R.R. (1940). A method of minimizing the sum of absolute values of deviations. *Ann. Math. Statist.* **11**, 301-310.
- [37] Turner, H.H. (1887). On Mr. Edgeworth's method of reducing observations relating to several quantities. *Phil. Mag.* (5th Series) **24**, 466-70.
- [38] Wagner, H.M. (1959). Linear programming techniques for regression analysis. *J. Am. Statist. Assoc.* **54**, 206.

# Abstracts

## Asymptotic properties of $L_1$ estimators in a multi-stage dose-response model: A Monte Carlo study

**Carmen D.S. André**, *University of São Paulo, Brazil*

**Subhash C. Narula**, *Virginia Commonwealth University, Richmond, USA*

*Abstract:* The single-stage and two-stage dose response models are frequently used in practical applications. The maximum likelihood and the least squares principles are often used to estimate the unknown parameters of the model. It has been shown that these methods are sensitive to outliers in the data. The minimum sum of absolute errors MSAE (or  $L_1$ ) criterion is more resistant to outliers than these popular procedures. However, at present not much is known about the statistical properties of the MSAE estimators of the parameters of the multistage dose-response model. In this paper, our objective is to study asymptotic properties and distribution of the MSAE estimators of the single-stage and two-stage dose-response models by simulation and to find the smallest sample size for which we may use the asymptotic distribution to draw statistical inferences about the parameters. We also give an approximate expression for the variance of these estimators when their asymptotic distribution follows a multinormal distribution.

## The $L^1$ -norm and interlaboratory tests

**P. L. Davies**, *University of Essen, Germany*

*Abstract:* The form of interlaboratory test we consider is that where each of  $I$  laboratories returns exactly one reading for each of  $J$  samples. Such a test may be described by the random effects model

$$X_{ij} = X_i + a_j + \epsilon_{ij}, \quad 1 \leq i \leq I, \quad 1 \leq j \leq J.$$

The  $X_i$  represent the laboratory effects, the  $a$  the sample contaminations and the  $\epsilon_{ij}$  the measurement errors. The problem is to identify outlying observations and outlying laboratories. As we have only one observation per cell it is commonly believed that it is not possible to detect outliers or, equivalently, non-additivity. As shown in Terbeck and Davies (1996) this is not correct and so called unconditionally identifiable outlier patterns may be found by the  $L^1$ - or an appropriate  $M$ -functional. The results



of Terbeck and Davies are improved in certain respects and then applied to the random effects model. The method is applied to a real data set considered by Lischer (1993) which is concerned with the determination of lead in sewage sludge.

## Multivariate $L_1$ mean

**Yadolah Dodge and Valentin Rousson,**

*University of Neuchâtel, Switzerland*

*Abstract:* The center of a univariate data set  $\{x_1, \dots, x_n\}$  can be defined as the point  $\mu$  that minimizes the norm of the vector of distances  $\mathbf{y}' = (|x_1 - \mu|, \dots, |x_n - \mu|)$ . As the median and the mean are the minimizers of respectively the  $L_1$ - and the  $L_2$ -norm of  $\mathbf{y}$ , they are two alternatives to describe the center of a univariate data set. The center  $\mu$  of a multivariate data set  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  can also be defined as minimizer of the norm of a vector of distances. In multivariate situations however, there are several kinds of distances. In this paper, we consider the vector of  $L_1$ -distances  $\mathbf{y}'_1 = (||\mathbf{x}_1 - \mu||_1, \dots, ||\mathbf{x}_n - \mu||_1)$  and the vector of  $L_2$ -distances  $\mathbf{y}'_2 = (||\mathbf{x}_1 - \mu||_2, \dots, ||\mathbf{x}_n - \mu||_2)$ . We define the  $L_1$ -median and the  $L_1$ -mean as the minimizers of respectively the  $L_1$ - and the  $L_2$ -norm of  $\mathbf{y}_1$ ; and then the  $L_2$ -median and the  $L_2$ -mean as the minimizers of respectively the  $L_1$ - and the  $L_2$ -norm of  $\mathbf{y}_2$ . In doing so, we obtain four alternatives to describe the center of a multivariate data set. While three of them have been already investigated in the statistical literature, the  $L_1$ -mean appears to be a new concept. Contrary to the  $L_1$ -median, the  $L_1$ -mean is proved to be unique in almost all situations. In order to compare these multivariate medians and means, we use the rule of the net advantage coefficient introduced by Stavig and Gibbons (1977). A simulation study shows that the  $L_1$ -mean performs well, especially for data sets drawn from bivariate Laplace distribution.

## The information for the direction of dependence in $L_1$ regression

**Yadolah Dodge,** *University of Neuchâtel, Switzerland*

**Joe Whittaker,** *Lancaster University, U.K.*

*Abstract:* An  $L_1$  regression model for a response variable  $X_2$  is to suppose that the conditional distribution of  $X_2$  given  $X_1$  is Laplace, and that the marginal distribution of the explanatory variable  $X_1$  is also Laplace.

We show that there is information to distinguish the direction of dependence  $X_1$  and  $X_2$ ; or equivalently to distinguish between the models in which  $X_1$  is dependent on  $X_2$ , and  $X_2$  is dependent on  $X_1$ . This is not true for  $L_2$  regression based on the Normal distribution.

## **Dimension choice for sliced inverse regression based on ranks**

**Louis Ferré**, *Université Paul Sabatier, Toulouse, France*

*Abstract:* Sliced Inverse Regression is a method for reducing the dimensionality in multivariate non parametric regression problems. While the selection of the dimensionality has been investigated for the original version, no solution has been proposed for Hsing and Carroll (1992) approach based on order statistics and associated concomitant variables. By using model selection approaches, we propose here two ways for selecting the dimensionality by estimating a loss function: first, a direct estimation is proposed and, then a Jack-Knifed estimate is investigated. Finally, the rank version is compared to classical SIR on a real life data set.

## **$L_1$ -norm and $L_2$ -norm methodology in cluster analysis**

**Allan D. Gordon**, *University of St Andrews, North Haugh, Scotland*

*Abstract:* An overview is provided of the use of  $L_1$ -norm and  $L_2$ -norm methodology in cluster analysis. Topics covered include dissimilarity measures, partitions, fuzzy classifications, hierarchical classifications, and consensus classifications.

## **Some issues in the applications of conditional quantile functions**

**Xuming He**, *University of Illinois, USA*

*Abstract:* Conditional quantile functions are useful in a variety of applications. Regression quantiles for linear models have been recently extended to semiparametric and nonparametric models. Further investigations are needed for both the statistical theory and computations. In this paper, I attempt to raise two questions that I believe are important to build a

solid foundation for the applications of quantile regression. They focus on nearly extreme quantiles and the problem of crossing in estimated quantile functions.

## **The median function on structured metric spaces**

**F. R. McMorris**, *University of Louisville, USA*

*Abstract:* When  $(X, d)$  is a finite metric space and  $\pi = (x_1, \dots, x_k) \in X^k$ , a median for  $\pi$  is a element  $x$  of  $X$  for which  $\sum_{i=1}^k d(x, x_i)$  is minimum. The function that returns the set of all medians for any tuple  $\pi$  is called the median function on  $X$ . A brief survey is given of some of the results concerning the median function, starting with an arbitrary metric space and finishing with the case where  $X$  is a set of hypergraphs and  $d$  is the metric based on the  $L_1$ -norm. A simplistic maximum likelihood interpretation for the median function is also given.

## **Least absolute value estimation of a linear functional model**

**Édina Shisue Miazaki**, *Statistika Consultoria, Campinas, Brazil*  
**Gabriela Stangenhau**, *Universidade de Brasília, Brazil*

*Abstract:* This paper presents two robust  $L_1$  based estimators for the parameters of a simple linear functional relationship (SLFR). The maximum likelihood estimation when the errors follow a double exponential distribution and the weighted  $L_1$  estimation are solved as non-linear optimization problems. The least median of squares estimates are proposed as starting values and the scale measures of the errors are based on the MAD. Both methods are resistant to outlying observations and the weighted  $L_1$  estimator is resistant to leverage points. Real examples illustrate the methods.

## **ANOVA - models: A Bayesian analysis**

**Wolfgang Polasek and Shuangzhe Liu**, *University of Basel, Switzerland*

*Abstract:* The 1-way and 2-way ANOVA are formulated as Bayesian linear models with conjugate prior distributions. The classical case is treated

as a special one using matrix generalized (g-) inverses leading to so-called  $OLS^-$  and  $OLS^+$  estimates of the rank deficient ANOVA model. The 2-way ANOVA model without interactions can also be estimated in a 2-step procedure.

## **Robustifying growth curve model estimation**

**Gabriela Stangenhau**, *Statistika Consultoria, Campinas, Brazil*  
**Elisete C. Quintaneiro Aubin**, *Universidade de São Paulo, Brazil*

*Abstract:* A robustified version of the parameter matrix estimators in the standard growth curve model obtained via the Potthoff-Roy transformation is presented. The asymptotic distribution of the robust estimators is derived and the estimation of their variance-covariance matrix is discussed.

## **Fitting $L_2$ norm classification models to complex data sets**

**Maurizio Vichi**, *University "G.DAnnunzio" di Chieti, Pescara, Italy*

*Abstract:* In this paper methodologies for fitting classification models (dendrograms and partitions) to two and three-way arrays of dissimilarities minimizing a  $L_2$  norm loss function are examined. A new algorithm for fitting several hierarchical classifications to quite large three-way arrays is also discussed.

## **Applications of mathematical programming in $L_1$ -estimation of nonlinear models**

**Jinde Wang**, *Nanjing University, China*

*Abstract:* In this paper we review the results in  $L_1$ -estimation of nonlinear models obtained by applying mathematical programming techniques. We describe briefly the ways to find the asymptotic distribution, the approximate representation and to treat dependent random error cases and inequality-constrained cases. With these results one can conclude that mathematical programming is a suitable tool for studying  $L_1$ -estimation problems.

## Some contributions to M-estimation in regression models

**Lincheng Zhao,**

*University of Science and Technology of China, Hefei, China*

*Abstract:* In this paper, we briefly survey some contributions to asymptotic theory on M-estimation in a linear model and least absolute deviations (LAD) estimation in a censored regression model (known as the Tobit model), as well as on the relevant test criteria in ANOVA in the above models. As a general approach on statistical data analysis, asymptotic theory of M-estimation in regression models has received extensive attention. In recent years, the author and some of his cooperators worked on this field and obtained some new results. In this paper we briefly introduce some of them and the related work in the literature. As a special case, the minimum  $L_1$ -norm ( $ML_1N$ ) estimation, also known as the least absolute deviations (LAD) estimation, plays an important role and is of special interest. Considering this point, we will pay much attention to them as well. Our topics concern the usual linear model and a censored regression model, known as the Tobit model in econometric research.

# Contributors

Adcock Christopher, School of Economic and Business Studies, Univ. of Westminster, 309 Regent Streest, W1R 8AL London, U.K.

Agullo Jose, Dept. Fundamentos del Analisis Economico, Universidad de Alicante, 03080 Alicante, SPAIN

Andre D.S. Carmen, Instituto de Matematica e Estatistica da USP, Rua do Matao 1010, C.P. 66281 - Agencia Cidade de Sao Paolo, CEP 05315-970 Sao Paulo-SP, BRAZIL

Arabie Phipps, Graduate School of Management, Rutgers University, 92, New Street, NJ 07102-1895 Newark, USA

Aubin Elisete, Instituto de Matematica e Estatistica da USP, Rua do Matao, 1010, CEP, 05508-900 Sao Paulo, SP, BRAZIL

Bassett Gilbert W., Dept. of Economics, Univ. Illinois-Chicago, 601 S. Morgan St, Rm 2103, 60607-7121 Chicago/ Illinois, USA

Brown Bruce M., Univesity of Tasmania-Math, P.O. Box 252 C, 7001 Hobart Tasmania, AUSTRALIA

Bura Efstathia, Dept. of Statistics, George Washington Univ., DC 20052 Washington, USA

Cadarso-Suarez Carmen, Dept de Estadistica e I.O., Seccion de Bioestadistica, Facultad de Medicina, Universidad de Santiago, La Coruna Santiago de Compostela, SPAIN

Carroll Douglas, Faculty of Management/ Marketing, Rutgers Univ., 81 New St, NJ 07102-1895 Newark, USA

Chakraborty Biman, Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, 203 B.T. Road, 700035 Calcutta, INDIA

Chaturvedi Anil, Technology Consultant, AT&T Labs, Room 5C-134, 600 Mountain Avenue, NJ 07974 Murray Hill, USA

Chaudhuri Probal, Theorical Stats & Math., Indian Stat. Inst., 8 Ramlal Agarwalla Lane, 700050 Calcutta, INDIA

Chepoi Victor, Laboratoire de Biomathmatiques, Faculté de Médecine, Université d'Aix Marseille II, 27 Bd Jean Moulin, 13385 Marseille Cedex 5, FRANCE

Chiaromonte Francesca, International Institute for Applied Systems Anal-

ysis, A-2361 Laxenburg, AUSTRIA

Cogneau Daniel, Faculté de Médecine, Laboratoire de Biomathématique, 27, Bd. Jean Moulin, 13385 Marseille, FRANCE

Davies Laurie, Univ. Gesamthochschule Essen, FB6 Maths, Universitätstrasse 3, 4300 Essen 1, GERMANY

de Una Alvarez Jacobo, Depto de Estadística e I.O., Campus Lagoas-Marcosende, Universidad de Vigo, 86200 Vigo, SPAIN

Dodge Yadolah, Groupe de Statistique, Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, SWITZERLAND

Farebrother Richard William, 11 Castle Road, Bayston Hill, Shrewsbury, SY3 ONF Shropshire, U.K.

Fernholz Luisa, Statistics Dept., Temple Univ., Speakman Hall, PA 19122 Philadelphia, USA

Ferré Louis, Laboratoire de Statistique, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, FRANCE

Fichet Bernard, Faculté de Médecine, Laboratoire de Biomathématique, 27, Bd. Jean Moulin, 13385 Marseille, FRANCE

Fitzenberger Bernd, Department of Economics, and Statistics, University of Konstanz, P.O. Box 5590 < D 139 >, D-78434 Konstanz, GERMANY

Fraiman Ricardo, Centro de Matematica, Universidad de la Republica-Uruguay, Eduardo Acevedo 1139, Montevideo, URUGUAY

Gordon Allan D., Mathematical Institute, University of St Andrews, North Haugh, St Andrews KY16 9ss, Fife, SCOTLAND

Gonzalez-Manteiga Wenceslao, Departamento de Estadística, Facultad de Matematicas, Universidad de Santiago de Compostela, La Coruna, SPAIN

Gutenbrunner Cornelius, Uni-Klinikum, Abteilung für Kinder und Jugendpsychiatrie, Hans-Sachs Strasse 6, 35033 Marburg, GERMANY

Hall Peter, Dept. of Statistics, A.N.U., G.P.O. Box 4, A.C.T. 2601 Canberra, AUSTRALIA

Hallin Marc, Institut de Statistique, Campus de la Plaine, Univ. Libre de Bruxelles, C.P. 210, B-1050 Bruxelles, BELGIUM

He Xuming, Dept. of Statistics, Univ. of Illinois-Champaign, 725 S Wright St, IL 61820-5714 Champaign, USA

Hettmansperger Tom, Pennsylvania State Univ., Dept. of Statistics, 317 Classroom Building, PA 16802-2111 University Park, USA

Huber, Peter J., LSt. Mathematik, Universität Bayreuth, Universitätsstrasse 30, 95440 Bayreuth, GERMANY

Hubert Lawrence, Dept. of Psychology (and Dept. of Statistics), University of Illinois-Champaign, 603 East Daniel Street, IL 61820-6232 Champaign, USA

Hubert Mia, Dept of Mathematics, Universitaire Inst. Antwerpen, Universiteitsplein 1, 2610 Wilrijk, BELGIUM

Hurst Simon R., Centre for Financial Mathematics, School of Mathematical Sciences, Australian National Univ., ACT 0200 Canberra, AUSTRALIA

Huskova Marie, Charles Univ., Dept. of Probability and Stat., Sokolovska 83, 18600 Prague 8, CZECH REPUBLIC

Jureckova Jana, Charles Univ., Dept. of Probability and Stat., Sokolovska 83, 18600 Prague 8, CZECH REPUBLIC

Kariya Takeaki, Institute of Economic Research, Hitotsubashi Univ. Kunitachi, 186 Tokyo, JAPAN

Klebanov L.B., Institute of Mathematical Geology, St Petersburg, RUSSIA

Knight Keith, Dept. of Statistics, Univ. of Toronto, 100 St-George Street, M5S 1A1 Toronto, CANADA

Koenker Roger, Dept. of Economics, Univ. of Illinois, 1206 S 6th St, IL 61820-6915 Champaign, USA

Liu Regina, Dept. of Statistics-Hill Cntr, Rutgers University, NJ 08855 Piscataway, USA

Liu Shuangzhe, Institute of Statistics and Econometrics, Holbeinstrasse 12, 4051 Basel, SWITZERLAND

McMorris F.R., Univ. of Louisville, Dept. of Mathematics, KY 402 92 Louisville, USA

McKean Joe, Dept. of Math & Statistics, Western Michigan Univ., MI 49 008 Kalamazoo, USA

Meade Nigel, The Management School, Imperial College, 53 Princes Gate, SW7 2PG London, U.K.



Meloche Jean, Dept. of Statistics, 2021 West Mall, Univ. of British Columbia, BC V6T 1W5 Vancouver, CANADA

Meulman Jacqueline J., Department of Data Theory, Leiden University, P.O. Box 9555, 2300 RB Leiden, THE NETHERLANDS

Miazaki Edina Shisue, C.P. 04393, 70919-970 Brasilia, D.F., BRAZIL

Mirkin Boris G., Division of Theoretical Bioinformatics (Abt 0815), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, GERMANY

Mizera Ivan, Department of Probability and Statistics, Comenius University, Mlynska dolina, 84215 Bratislava, SLOVAKIA

Moreno Elias, Department of Statistics, Faculty of Science, University of Granada, 18071 Granada, SPAIN

Morgenthaler Stephan, EPFL, Dept. of Mathematics, 1015 Lausanne, SWITZERLAND

Möttönen Jyrki, Department of Mathematical Sciences, University of Oulu, 90570 Oulu, FINLAND

Müller Christine, Freie Universität Berlin, Fachbereich Math. & Informatik, WE1, Arnimalle 2-6, 14195 Berlin, GERMANY

Narula Subash C., School of Business, Virginia Commonwealth Univ., 1015 Floyd Avenue, VA 23284-4000 Richmond, USA

Nolan Deborah A., Dept. of Statistics, 367 Evans Hall, Univ. of California, CA 94720-3860 Berkeley, USA

Nyquist Hans, Swedish Univ. of Agriculture, Dept. of Biometry, 901 83 Umea, SWEDEN

Oja Hannu, Dept. of Applied Maths and Statistics, Linnanmaa, Univ. Oulu, 90570 Oulu 57, FINLAND

Platen Eckhard, Centre for Financial Mathematics, School of Mathematical Sciences, The Australia National University, 0200 Canberra ACT, AUSTRALIA

Polasek Wolfgang, Institute of Statistics and Econometrics, Holbeinstrasse 12, 4051 Basel, SWITZERLAND

Portnoy Steve, Dept. of Statistics, Univ. of Illinois-Champaign, 725 S Wright St, IL 61820-5714 Champaign, USA

Raimondo Marc, Australian National University, Department of statis-

tics, G.P.O. Box 4, A.C.T. 2601 Canberra, AUSTRALIA

Rousseeuw Peter, Dept. of Mathematics, Universitaire Inst. Antwerpen, Universiteitsplein 1, 2610 Wilrijk, BELGIUM

Rousson Valentin, Groupe de Statistique, Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, SWITZERLAND

Sheather Simon, Australian Graduate School of Managment, University of New South Wales, 2033 Kensington NSW, AUSTRALIA

Stangenhau Gabriela, Rua Americo de Campos 133, 13083-040 Campinas SP, BRAZIL

Stute Winfried, Math. Institut, Univ. Giessen, Arndtastr. 2, 35392 Giessen, GERMANY

Terui Nobuhiko, Tinbergen Institute, Rotterdam, Burg Oudlaan 50, PA 3062 Rotterdam, THE NETHERLANDS

Vichi Maurizio, Universit "G. D'Annunzio" di chieti, Dept. of Metodi Quantitativi e Teoria Economica, Viale Pindaro 42, 65127 Pescara, ITALY

Wang Jinde, Dept. of Mathematics, Nanjing Univ., 22, Hankou Road, 210008 Nanjing, CHINA

Wetzel Nate, Dept. of Mathematical Sciences, Binghamton Univ., NY 13902-6000 Binghamton, USA

Whittaker Joe, Dept of Mathematics, Lancaster University, LA1 4YF Lancaster, U.K.

Zhao Lincheng, Department of Mathematics, Univ. of Science & Technology, HEFEI, 230026 Anhui, CHINA

Zwanzig Silvelyn, Institut für Mathematische Stochastik, Universität Hamburg, Bundesstrasse 55, 20146 Hamburg, GERMANY



## PROLOGUE

Me: *Which way must I follow?*

The Sage: *If you are a real pilgrim, you will achieve  
the journey whichever way you go.*

Sohravardi: *Persian Philosopher (1155-1191)*



# Measuring the performance of boundary-estimation methods

**Peter Hall**

*Australian National University, Canberra, Australia and  
CSIRO Mathematical and Information Sciences, Sydney, Australia*

**Marc Raimondo**

*Australian National University, Canberra, Australia*

*Abstract:* The problem of local linear approximation to a curved boundary using gridded data is closely connected to both curve estimation methods in statistics and rational approximation in number theory. The problem is ill-posed, in the sense that orders of approximation at arbitrarily close points can be very different. This may be interpreted as a consequence of the problem's number-theoretic aspects, since irrational numbers with arbitrarily slowly convergent rational approximations are distributed in dense sets. On the other hand, by measuring performance in a "statistically average" way which excludes most of the pathologies, we may deduce useful results about optimal orders of approximation. In this respect, among others, statistical approaches to the problem are important. For example, measures of performance based on the  $L^1$  norm are more appropriate than those founded on  $L^p$  norms for  $p > 1$ . The paper will describe these viewpoints, and outline the way in which they may be combined to produce a cohesive theory of curve estimation from gridded data. We shall start with the relatively simple case of approximation to a simple linear boundary, where data are observed without noise, and progress through an analysis of the number-theoretic connections, concluding with results in the context of stochastic or curved boundaries observed with noise.

*Key words:* Curve estimation, edge, gradient, grid, integral metric, irrational number, nonparametric, rational number, slope, vertex.

AMS subject classification: Primary 60G35; secondary 62G20.

## 1 Defining a straight-line boundary

Imagine placing a straight line across a square lattice in the plane, thereby dividing the plane into two parts. Assuming that the line is not vertical, colour black those lattice vertices above the line and white the vertices below, with a third colour (red, say) for any vertices that lie on the line. Now remove the line, and attempt to reconstruct it from the pattern of vertex colours. This is a theoretical idealisation of a range of practical boundary estimation problems, where a curve representing the boundary between two areas of different colour is to be estimated from pixel data.

Even a brief consideration of this problem shows that its solution depends critically on the nature of the slope of the line. For example, if the slope is rational and if the line passes through some vertex, then the line necessarily passes through an infinite number of vertices. In this case, if we were able to observe the vertex colour pattern in a large enough region of the plane, we would see that there are at least two red vertices there, and from them we could trivially deduce the equation of the line. Then, we would know the line exactly.

On the other hand, if the line has rational slope but does not pass through any vertex, it cannot be determined exactly even if we know the colour of every vertex in the plane. This is perhaps most easily seen if the line,  $\mathcal{L}$  say, is parallel to one of the axes of the square lattice. In that case there exists an infinite strip in the plane, with its sides parallel to the line and its width equal to the edge width of the lattice, such that any straight line contained wholly within the strip produces exactly the same vertex colour pattern as  $\mathcal{L}$ .

A similar situation arises for any line with rational gradient, where the intercept is chosen so that the line does not pass through any vertex. In such cases, while the gradient may be determined exactly from vertex colour data within the whole plane, the intercept will remain unknown beyond the fact that it lies within a certain nondegenerate interval — except when the line passes through a vertex. So, in the case of a line with rational slope we know either everything or nothing: either we can compute the line exactly from a finite amount of vertex colour data (when the line passes through a vertex) or we cannot compute it exactly even from an infinite amount of data (if it does not pass through any vertex).

The situation is quite different if the line has irrational slope. There, if the colour pattern is observed within an increasingly large region  $\mathcal{R}$ , say an  $n \times n$  section of the lattice centred roughly on the line, then an approximation to the line may be constructed using only the colour pattern within  $\mathcal{R}$ . As  $\mathcal{R}$  expands, the accuracy with which the line may be approximated

increases. More explicitly, we may compute an approximation  $\hat{\mathcal{L}} = \hat{\mathcal{L}}(\mathcal{P})$  to  $\mathcal{L}$ , using only the vertex colour pattern  $\mathcal{P}$  within  $\mathcal{R}$ , such that the Hausdorff distance between  $\hat{\mathcal{L}} \cap \mathcal{R}$  and  $\mathcal{L} \cap \mathcal{R}$  converges to zero as  $\mathcal{R}$  increases.

In the case of irrational slope the rate of convergence of a good approximation  $\hat{\mathcal{L}}$  depends intimately on the nature of the irrational slope. It depends hardly at all on whether  $\mathcal{L}$  intersects a lattice vertex; this influences only the constant multiple of the optimal rate of convergence of  $\hat{\mathcal{L}}$  to  $\mathcal{L}$ , not the rate itself. Thus, the problem of approximating straight-line boundaries is starkly ill-posed, since nearby slopes can produce very different convergence rates along infinite subsequences.

In Section 2 we shall treat examples of classes of irrational numbers, which capture the spirit of the boundary approximation problem and its solution. Section 3 will employ the examples to motivate development of more general boundary approximation problems, and will discuss ways in which the problems might be tackled. Section 4 will briefly survey the number-theoretic background to the methods. Later sections will develop theories for curved boundary estimation using local linear methods, borrowing ideas that are now well understood in more traditional statistical settings. For the latter, the reader is referred to Wand and Jones (1995, Chapter 5) and Fan and Gijbels (1996).

In Sections 1–5 we shall always assume that the lattice is fixed; without loss of generality it has its vertices at integer pairs  $(i, j)$  in the Cartesian plane, so that its edge width (the width of the side of the smallest square face of the lattice) is 1. In later sections we shall sometimes consider lattices of increasing fineness, so as to model the physical problem of approximating a curved boundary on a fine pixel grid. Technical details behind our arguments may be found in Hall and Raimondo (1996a,b).

While we shall concentrate on the case of a square lattice, for definiteness, the results that we shall describe are valid for any regular lattice that has the property that it contains a square lattice and is contained within the union of a finite number of square lattices. Thus, our results are available for lattices whose faces are hexagons or triangles. Lattices of the latter type are used in practice in J.P. Serra’s image analyser. When considering an “ $n \times n$ ” portion of a general lattice we interpret  $n$  as the square root of the number of vertices within a finite, square subset of the lattice.

## 2 Classes of irrational numbers

The irrational numbers with which many of us are most familiar are the so-called “quadratic irrationals”, defined as the set of real numbers that may be expressed as solutions of quadratic equations with rational coefficients



(or, without loss of generality, integer coefficients). These are a subset of the class of so-called *periodic irrationals*, and also of the larger class of *badly approximable* irrationals, which we shall discuss in Section 5. Straight-line boundaries with slope coming from one of these classes have special properties with respect to the boundary approximation problem. Indeed, in such cases the optimal rate of convergence (in the sense of the Hausdorff metric) of approximations based on vertex colours within an  $n \times n$  subset of the lattice, is asymptotic to a constant multiple of  $n^{-1}$ .

The set of *algebraic irrationals* is larger than the class of quadratic irrationals, and is defined as the set of all real numbers that may be expressed as solutions to polynomial equations with rational coefficients. However, the most accurate available estimate of the rate of convergence in the linear boundary approximation problem for boundaries with slope equal to an algebraic irrational, is only the upper bound of  $O(n^{-1+\epsilon})$  for all  $\epsilon > 0$ . Not even logarithmic refinements, such as  $O(n^{-1} \log n)$ , are available. The upper bound  $O(n^{-1+\epsilon})$  is a corollary of deep number-theoretic work of Roth (1955), who determined the exact exponent in the Thue-Siegel inequality and for which work he was awarded the Fields medal in 1958.

Roth's result is virtually equivalent to the upper bound  $O(n^{-1+\epsilon})$ , for all  $\epsilon > 0$ , in our approximation problem. If we could improve on that rate then we could refine Roth's Theorem, as it is known. And of course, even if we could refine Roth's result, we would still have only scratched the surface as far as solving our problem goes, since the great majority of irrational numbers are not algebraic. Indeed, since the number of rationals is countable then the number of polynomial equations of degree  $p$  with rational coefficients is countable. Therefore, the number of solutions of such equations, for arbitrary  $p$ , is countable. Hence, the number of algebraic irrationals is countable, whereas the number of irrational numbers is uncountably infinite. By focusing only on algebraic irrationals we would be missing the great majority of irrational numbers.

It might be thought that because the algebraic irrationals are dense in the real line, they provide a good guide to the sort of behaviour that will be experienced when the slope of the line is a non-algebraic irrational. While this is true from some viewpoints, the argument has limitations. Indeed, irrationals that are not algebraic, and produce particularly pathological convergence rates in our line approximation problem, also comprise a dense subset of the real line. To elucidate this point we mention that if  $\alpha_1, \alpha_2, \dots$  and  $\beta_1, \beta_2, \dots$  are any two sequences of positive numbers converging to zero, then there exists a dense set of irrational numbers such that, whenever the slope of the linear boundary is in this set, the optimal rate of convergence in our approximation problem on an  $n \times n$  grid is bounded above by  $\alpha_n$  along

one subsequence of values of  $n$ , and bounded below by  $\beta_n$  along another subsequence.

This is perhaps not a major issue if we confine attention to exactly linear boundaries — we may simply exclude such pathological irrational numbers from contention as possible gradients. However, in the problem of local linear estimation of a curved boundary the range of values of the gradient is an interval, and so includes representatives from any set which is dense in the real line. This fact, and the properties of irrational numbers noted above, make it clear that one must take care when defining boundary-estimation problems, to avoid becoming side-tracked by relatively unimportant cases.

### 3 Defining boundary-estimation problems

We need to pose boundary-estimation problems in such a way that we can deduce relatively simple principles behind rate-of-convergence properties. For that, we need some way of averaging over all possible choices of irrational gradients, so that the central issues in the problem will not be lost in through consideration of pathological special cases. There are at least two ways of doing this.

First, we might allow the slope of the boundary to be a random variable, and devote our discussion to its “average” properties. This is feasible for either straight or curved boundaries. If the boundary is linear then we may apply a random rotation to it, and more generally we may regard the boundary as a realization of a random curve whose equation is represented by  $y = G(x)$ , where  $G$  is a random, smooth function. Alternatively, we may choose to treat the boundary as fixed and curved, but estimate it at a randomly chosen point. Under such models we do not need to be too prescriptive about the type of averaging, since the more radical of the pathological cases described in the previous section arise only for sets of irrational numbers having measure zero. Therefore, if the random boundary, or the random point at which we estimate a fixed, curved boundary, is distributed in the continuum, then, by confining attention to almost sure properties we avoid all but reasonably regular cases. We shall outline this approach in Section 6.

Alternatively, in the case of a curved boundary we may average approximations in some way, for example by considering them in an integral metric. It turns out that the  $L^1$  metric is more appropriate for this purpose than an  $L^p$  metric for  $p > 1$ , since it is relatively resistant to large deviations in the approximation error. (In view of the properties described in Section 2, it comes as no surprise to learn that the approximation error can change dramatically as we move from one point to another along a curved

boundary.)

While the integral metric approach is attractive, not least because it is well established in the context of nonparametric curve estimation, it does require care. For example, if the boundary is linear then we are still faced with the ill-posed problem of the effect of rational-versus-irrational slope. The remedy is to avoid linear boundaries altogether. Now, one way of characterising a nonlinear boundary is to insist that its second derivative never vanish. Therefore, in Section 7 we shall study the  $L^1$  performance of local linear approximations to twice-differentiable boundaries that do not have any points of inflexion.

## 4 Rational approximation by continued-fraction expansion

In order to appreciate the methods and results for general boundary-estimation problems it is necessary to understand the main elements of the theory of rational approximation by continued fractions. We shall survey them here, referring the reader to Leveque (1956, Chapter 9) and Khintchine (1963) for more detailed discussion. Section 5 will make the connections to boundary estimation explicit.

A non-integer real number  $u$  may be uniquely expressed as a continued fraction,

$$u = [a_0; a_1, a_2, \dots] = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \dots}}},$$

where  $a_0$  is an integer and  $a_1, a_2, \dots$  are strictly positive integers, called the *partial denominators* of  $u$ . The continued fraction expansion terminates if and only if  $u$  is rational. Up to the termination point (in the case of rational  $u$ ), or for all  $n$  (if  $u$  is irrational), the *convergents* of  $u$  are defined to be the numbers

$$\frac{p_0}{q_0} = a_0, \quad \frac{p_1}{q_1} = a_0 + \frac{1}{a_1}, \quad \frac{p_2}{q_2} = a_0 + \frac{1}{a_1 + \frac{1}{a_2}}, \quad \dots,$$

where  $p_n$  and  $q_n$  are relatively prime integers. The  $q_n$ 's are strictly positive and form a strictly increasing sequence. By definition,  $p_n/q_n$  converges to  $u$  as  $n \rightarrow \infty$ . The sequence of odd-indexed convergents is decreasing, and the sequence of even-indexed convergents increases.

If  $u$  is irrational then the convergents provide a sequence of rational approximations to  $u$ , often referred to as “continued fraction approximations”. The approximations are optimal in the sense that

$$\inf_{p, 1 \leq q \leq q_n} |u - (p/q)| = |u - (p_n/q_n)|. \quad (1)$$

They also satisfy

$$\{q_n(q_n + q_{n+1})\}^{-1} < |u - (p_n/q_n)| < (q_n q_{n+1})^{-1} \quad (2)$$

and

$$\begin{aligned} &\text{if } p \text{ and } q \text{ are relatively prime, and } |u - (p/q)| \\ &< (2q^2)^{-1}, \text{ then } p/q \text{ is a convergent of } u \end{aligned} \quad (3)$$

The quality of approximations by continued fractions is determined mostly by properties of large elements of the sequence  $\{a_n\}$ , or equivalently by large values of  $q_{n+1}/q_n$ , since it may be shown that  $q_{n+1}/q_n \asymp a_n$  (meaning that the ratio of the left- and right-hand sides is bounded away from zero and infinity).

## 5 Relationship between convergents and rates of approximation to linear boundaries

The importance of continued fraction expansions to the problem of approximating linear boundaries, as defined in Section 1, is that the optimal rate of approximation (in the Hausdorff metric) to a line  $\mathcal{L}$  with irrational gradient  $u$ , using vertex colour data observed on an  $n \times n$  grid, is essentially equivalent to  $n$  times the optimal rate at which we can approximate  $u$  by a rational number  $p/q$  with  $q$  not exceeding  $n$ . In view of properties (1)–(3) the latter rate is the order of

$$\{q_{k(n)}(u) q_{k(n+1)}(u)\}^{-1},$$

where  $k(n) = k(n, u)$  denotes the smallest  $k$  such that  $q_k(u) \leq n$ . Call these results (R). A relationship between rational approximations and lines on square lattices is also expressed by Klein diagrams; see for example Klein (1907).

To illustrate the importance of the connection between convergents and boundary approximation we consider a simple example. The real number  $u$  is said to be *badly approximable* (or BA, for short) if  $\sup_n a_n(u) < \infty$ . The set of all BA numbers in the interval  $[0, 1]$  has cardinality equal to that of the continuum (see e.g. Schmidt 1980, p. 23), but is of measure zero (e.g. Khintchine 1963, p. 69). All quadratic irrationals are BA, since for them the sequence  $\{a_n\}$  is eventually periodic. However, not all algebraic irrationals are BA. In view of the asymptotic equivalence of the sequences  $a_n$  and  $q_{n+1}/q_n$ , and the fact that  $q_n$  is increasing,  $u$  is BA if and only if  $(q_{k(n)} q_{k(n+1)})^{-1}$  is bounded between two constant multiples of  $n^{-2}$ .

From this result and (R) we see that an irrational number  $u$  is BA if and only if the optimal rate (in the Hausdorff metric) at which a line

with gradient  $u$  may be approximated from vertex colour data in an  $n \times n$  section of the lattice, is  $n^{-1}$ . As a corollary, the optimal rate is  $n^{-1}$  when the gradient of  $\mathcal{L}$  is a quadratic irrational.

## 6 A stochastic number-theoretic view

Khinchine (1963), describing and developing work dating from the 1930's (see e.g. Khinchine, 1935; Lévy, 1937), gave a concise account of rates of rational approximation to irrational numbers when the latter are chosen randomly with respect to Lebesgue measure. In view of the equivalence between problems of rational approximation and boundary approximation noted in Section 5, we may apply Khinchine's results to our line estimation problem.

To pose that problem in a stochastic setting we assume that the linear boundary is placed into the plane according to a random mechanism. For our purposes the mechanism may be defined very generally; we need only ask that the distribution of slope, conditional on the line's intercept with any given axis, be continuous. This reflects the fact that, when the line has irrational gradient — which it will enjoy with probability 1 if the gradient has a continuous distribution — it is immaterial from the viewpoint of rates of approximation whether the line passes through a vertex.

It is known, for example from Theorem 30 of Khinchine (1963), that if  $\psi(n) = n^{-1}L(n)$  for a positive, slowly varying function  $L$  then, for almost all real numbers  $u$  (with respect to Lebesgue measure),  $\psi(n) q_{n+1}(u)/q_n(u) = O(1)$  if and only if

$$\sum_{n=1}^{\infty} \psi(n) < \infty; \quad (4)$$

and from Lévy (1937, p. 320) that  $n^{-1} \log q_n(u) \rightarrow \pi^2/(12 \log 2)$  as  $n \rightarrow \infty$ . Hence, for almost all  $u$ ,  $\psi\{\log q_n(u)\} q_{n+1}(u)/q_n(u) = O(1)$  if and only if (4) holds.

This result, and the relationship between rational approximation and linear boundary approximation discussed in Section 5, may be used to show that if the boundary is stochastic in the sense defined in the previous section, then with probability one the optimal rate of approximation to the boundary, in the Hausdorff metric restricted to a region containing an  $n \times n$  grid and using data from that grid, equals

$$O\{n^{-1}(\log n) L(\log n)^{-1}\} \quad (5)$$

if and only if (4) holds. Similarly, it may be proved that with probability one (4) is equivalent to asking that the optimal rate of approximation be

no better than

$$O\{(n \log n)^{-1} L(\log n)\} \quad (6)$$

along any subsequence.

The bound at (5) implies that if the line is placed into the plane at random according to the regime suggested above, then with probability 1 it can be approximated at rate

$$O\{n^{-1}(\log n)(\log \log n)(\log \log \log n)^{1+\epsilon}\}, \quad (7)$$

for all  $\epsilon > 0$ , from vertex colour data within an  $n \times n$  region of the square lattice; and the bound at (6) shows that the optimal convergence rate is no better than

$$O\{n^{-1}(\log n)^{-1}(\log \log n)^{-1}(\log \log \log n)^{-1-\epsilon}\},$$

along some subsequence. Moreover, these results are false if  $\epsilon$  is replaced by 0.

## 7 Approximations to curved boundaries

The results derived in Section 6 may be readily extended to the case of local linear approximations to smooth curves on a square lattice. There it is convenient to introduce the concept of a grid of increasing fineness, so as to develop a theory for curve estimation using increasing amounts of information. Rather than assume that the lattice has fixed edge width we suppose it has edge width  $n^{-1}$ . For example, we might suppose that its vertices are at points  $(n^{-1}i, n^{-1}j)$ , where  $i, j$  range over the set of all integers.

Replace  $\mathcal{L}$  by a smooth curve  $\mathcal{C}$ , for example given by the equation  $y = g(x)$ . As before, colour black the vertices above  $\mathcal{C}$  and white the vertices below, and consider constructing a local linear approximation to  $g$  at  $x$  by employing the colours of all vertices that lie within the strip  $\mathcal{S} = \mathcal{S}(x)$  defined by  $\{(t, y) : x - h < t < x + h \text{ and } -\infty < y < \infty\}$ . Here,  $h$  plays the role of bandwidth in more traditional curve estimation problems, and the asymptotics involve  $h = h(n)$  converging to zero as  $n \rightarrow \infty$ , in such a manner that  $nh \rightarrow \infty$ .

We may define the local linear approximant,  $\bar{g}(x)$ , at  $x$  to be any straight-line segment that agrees with the vertex colour pattern within  $\mathcal{S}(x)$ ; or any segment that has least number of disagreements, if no segment agrees completely. There are two sources of error in this approximation. First, there is a degree of bias, or systematic error, due to the fact that the part of  $\mathcal{C}$  that lies within  $\mathcal{S}$  is not exactly a straight line. As in more

familiar, second-order nonparametric curve estimation problems, the bias is  $O(h^2)$  as  $h \rightarrow 0$  if  $g$  has two bounded derivatives. Secondly, there is approximation error arising from the fact that our only information about  $g$  is in the form of vertex colours. If the problem of estimating  $g(x)$  has a random component, for example if  $x$  is taken to be a random variable with an absolutely continuous distribution, then the results developed in Section 6 for the case of a random line may be applied directly to the setting of approximating a random curve by a line segment within  $\mathcal{S}$ .

In particular, formula (7) may be used to bound the second type of approximation error, provided we replace  $n$  by  $nh$  and allow  $nh$  to increase without limit. Then, assuming that  $\log(nh)$  increases like  $\log n$ , which will certainly be the case for optimal choice of  $h$ , we see that the second type of error is bounded above by

$$O\{(n^2h)^{-1}(\log n)^{1+\epsilon}\}. \quad (8)$$

Optimising the over-all convergence rate involves balancing systematic and non-systematic sources of error; that is, choosing  $h$  so that the bias term, of order  $h^2$ , is of the same size as the quantity at (8). This means taking  $h$  to be of size  $\{n^{-2}(\log n)^{1+\epsilon}\}^{1/3}$ , which gives a convergence rate of  $O[\{n^{-2}(\log n)^{1+\epsilon}\}^{2/3}]$ . The rate  $n^{-4/3}$ , multiplied by a positive power of  $(\log n)^{-1}$ , may be shown to be a minimax lower bound in this problem. In related work, Korostelev and Tsybakov (1993) have shown that the rate  $n^{-4/3}$  is minimax optimal in the case of certain random grids. Thus, the local linear approximation  $\bar{g}$  is within at most a logarithmic factor of achieving the optimal rate.

## 8 An $L_1$ view of boundary approximation

In the account of boundary approximation just above, we incorporated an element of randomness in order to remove the ill-posed nature of the problem. Without that randomness, the pointwise properties of rates of convergence defy simple description. Alternatively, we may address global rates of convergence in an  $L^p$  metric. We know from the work in earlier sections that, while there are many pathological cases where convergence rates are arbitrarily poor (along subsequences), and while such cases arise at points forming a dense set, they have measure zero. Hence, we are entitled to expect that they will not loom excessively large in an  $L^p$  measure of performance. Since the case  $p = 1$  puts least emphasis on very large errors then it is potentially the most useful.

Let  $\mathcal{C}$  have equation  $y = g(x)$ , and suppose we observe the vertex colour pattern at all vertices  $(in^{-1}, jn^{-1})$  for integers  $i, j$  with  $0 \leq i \leq n$  and

$-\infty < j < \infty$ . (Thus, we adopt the “increasingly fine grid” model suggested in Section 7.) Construct the local linear approximation proposed in Section 7, so that for each  $x$  in a compact interval  $\mathcal{I}$  (which we take without loss of generality to be  $[0, 1]$ ) we have an approximation  $\bar{g}(x)$  to  $g(x)$  using vertex colour data within the strip  $\mathcal{S}(x)$ . Employing property (2) of convergents we may prove that if  $g$  has two bounded derivatives then, for absolute constants  $A_1$ ,  $A_2$  and  $A_3$ ,

$$|\bar{g}(x) - g(x)| < A_1 h \{q_{N(u)}(u) q_{N(u)+1}(u)\}^{-1} + B h^2, \quad (9)$$

for all  $x \in \mathcal{I}$ ,  $0 < h < \frac{1}{2}$  and  $nh > A_2$ , where  $u = g'(x)$ ,  $N = N(u)$  is the largest integer such that  $q_N(u) \leq A_3 nh$ , and  $B = \sup_{\mathcal{I}} |g''|$ . Here,  $q_n(u)$  is the denominator of the  $n$ 'th convergent,  $p_n(u)/q_n(u)$ ; see Section 4 for a definition.

The second term on the right-hand side of (9) derives from bias, or equivalently from the systematic error induced by approximating a nonlinear curve by a short but linear segment. The first term results from the limited information available about  $g$ , in the form of vertex colours. That term can be arbitrarily large, owing to the sort of pathology noted at the end of Section 2. However, provided  $g''$  is bounded away from zero the integral average of the first term is generally reasonable in size. In fact, it may be shown that if  $h = h(n) \rightarrow 0$  in such a manner that  $(nh)^{-2}(\log n)^2 \rightarrow 0$ , then

$$\int_{\mathcal{I}} |\bar{g}(x) - g(x)| dx = O\{(n^2 h)^{-1}(\log n)^2 + h^2\}, \quad (10)$$

uniformly in functions  $g$  for which, for some  $C > 1$ ,

$$C^{-1} \leq \inf_{x \in \mathcal{I}} |g''(x)| \leq \sup_{x \in \mathcal{I}} |g''(x)| \leq C. \quad (11)$$

The lower bound in (11) ensures that  $g$  is not too much like a straight line.

Choosing  $h$  of size  $(n^{-1} \log n)^{2/3}$  in (10) we obtain a rate of approximation, in the  $L^1$  metric, of  $O\{(n^{-1} \log n)^{4/3}\}$  uniformly in functions satisfying (11). Again, this convergence rate is close to the optimum of  $n^{-4/3}$ ; see Section 7.

In principle, similar results may be derived for rates of approximation in  $L^p$  metrics, where  $p > 1$ . However, those rates are inferior to the  $L^1$  rate by a polynomial order of magnitude. The reason is that, for  $p > 1$ , the  $L^p$  metric gives greater weight to larger values of the error  $|\bar{g}(x) - g(x)|$ .

To better appreciate the nature of this problem, observe from (9) that we have the bound

$$|\bar{g}(x) - g(x)| < A_1 A_3^{-2} h (nh)^{-2} \{q_{N(u)+1}(u)/q_{N(u)}(u)\} + B h^2,$$



which is potentially the key to deriving formulae such as (10). However, the ratio  $Q_n(u) = q_{N(u)+1}(u)/q_{N(u)}(u)$  is very unstable. Bear in mind that, when finding the integral average of the right-hand side, we are in effect taking  $U$  to be a random variable with the Uniform distribution on  $\mathcal{I}$ , and (in the case of the  $L^p$  metric) asking that  $E\{Q_n^p(U)\}$  be bounded. Now, it may be shown that the process  $\{Q_n(U), n \geq 1\}$ , is Markovian, and that (while the process is itself not stationary) it has a stationary limit distribution. Therefore,  $Q_n(U) = O_p(1)$  as  $n \rightarrow \infty$ . However, the stationary distribution does not have any finite moments, and in fact  $E\{Q_n^p(U)\} = \infty$  for all  $n$  and all  $p \geq 1$ . The term  $(\log n)^2$  on the right-hand side of (10) is the result of taking a more subtle approach to this problem, necessary even in the case  $p = 1$ .

## 9 Estimating boundaries observed with noise

The noiseless model introduced in Section 7 may be written in the form

$$Y(i/n, j/n) = I\{j/n \leq g(i/n)\},$$

where  $I(\cdot)$  is an indicator function,  $Y(i/n, j/n)$  denotes the colour of the vertex at  $(i/n, j/n)$  (white is represented by 1 and black by 0), and the equation  $y = g(x)$  represents the boundary  $\mathcal{C}$ . In practice, due to a combination of systematic and stochastic errors, the colour of each vertex may be more appropriately represented by a number between  $-\infty$  and  $\infty$ . In particular, we may write

$$Y(i/n, j/n) = f(i/n, j/n) + \epsilon_{ij},$$

where  $f(\cdot, \cdot)$  is a function with a fault-type discontinuity along the curve  $y = g(x)$ , and the independent and identically distributed stochastic errors  $\epsilon_{ij}$  have zero mean.

It will be assumed that  $f$  admits the representation

$$f(x, y) = f_1(x, y) + f_2(x, y) I\{y \leq g(x)\},$$

where  $f_1$  and  $f_2$  each have two uniformly bounded derivatives of all types, and  $f_2$  is bounded away from zero. We suppose that  $g$  and its first two derivatives are bounded on  $\mathcal{I}$ . Several different analogues of the local linear estimators in Section 7 are possible; examples include versions based on least squares and on wavelets. We consider here only the former. It amounts to first computing a preliminary approximation,  $\tilde{g}$ , and then refining it using local linear smoothing within a window. We shall consider a particularly simple preliminary estimator, based on kernel methods, as follows.

Suppose we wish to estimate  $g$  at  $x \in \mathcal{I}$ . Write  $i_n$  for the integer nearest to  $nx$ , let  $K$  be a nonnegative, compactly supported, continuously differentiable function, let  $h_1$  equal a constant multiple of  $n^{-2/3}$ , and put

$$T(j) = (nh_1^2)^{-1} \sum_k K'\{(j-k)/(nh_1)\} Y(i_n/n, j/n),$$

which is a statistical approximation to the first derivative of  $f(i_n/n, \cdot)$  at  $j/n$ . Let  $\hat{j}$  denote a value which produces a global maximum of  $|T|$  in the range  $C_1 n \leq j \leq C_2 n$ . Our preliminary estimator of  $g(x)$  is  $\tilde{g}(x) = \hat{j}/n$ .

Next we define an improved estimator. Let  $\mathcal{W}$  be a square window of side length  $h = h(n)$ , with its centre at  $(i_n/n, \hat{j}/n)$  and, for the sake of definiteness, its axes aligned with those of the grid. Temporarily make the assumption that within  $\mathcal{W}$ ,  $f$  assumes a constant value on either side of a line  $\mathcal{L}$ . We fit  $\mathcal{L}$  by least squares in the class  $\mathcal{M}(C, \mathcal{W})$  of all lines  $\mathcal{L}$  that divide  $\mathcal{W}$  into two sets of vertices of which the larger has no more than  $C$  times the number in the smaller (where  $C > 1$  is arbitrary but fixed). Specifically, let  $\mathcal{I}_1$  [respectively,  $\mathcal{I}_2$ ] denote the set of vertex coordinates  $w = (i/n, j/n)$  in  $\mathcal{W}$  that lie above [below]  $\mathcal{L}$ , let  $\sum^{(i)}$  denote the sum of  $Y(w)$  over all  $w \in \mathcal{I}_i$ , let  $\bar{Y}_i$  be the corresponding mean, and put

$$S(\mathcal{L}) = \sum_{i=1}^2 \sum^{(i)} \{Y(w) - \bar{Y}_i\}^2.$$

Write  $\hat{\mathcal{L}}$  for a line that minimizes  $S(\mathcal{L})$  among all straight lines in  $\mathcal{M}(C, \mathcal{W})$  that do not pass through any vertices. (The minimum is of course not uniquely attained, and any measurable approach to breaking ties is allowed.) Write  $\hat{g}(x)$  for the ordinate of the point on  $\hat{\mathcal{L}}$  with abscissa  $x$ .

Provided the distribution of the errors  $\epsilon_{ij}$  has sufficiently light tails, it may be proved that  $\hat{g}$  has properties similar to those ascribed to the local linear estimator  $\bar{g}$  in the no-noise case in Sections 7 and 8. For example, let us assume that the moment generating function of the error distribution exists and is finite in a neighbourhood of the origin. If  $g''$  is bounded and  $X$  is a continuous random variable (stochastically independent of the errors  $\epsilon_{ij}$ ), then it may be shown that, for suitable choice of  $h$ , with probability one  $\hat{g}(X)$  converges to  $g(X)$  at rate  $O\{n^{-2}(\log n)^{1+\epsilon}\}^{2/3}$  (compare Section 7); and if  $|g''|$  is bounded away from both zero and infinity then, again for appropriate  $h$ ,  $\hat{g}$  converges to  $g$  in  $L^1$  at rate  $O\{(n^{-1} \log n)^{4/3}\}$ .

## References

- [1] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.

- [2] Hall, P. and Raimondo, M. (1996a). Approximating a line thrown at random onto a grid. *Ann. Appl. Probab.* To appear.
- [3] Hall, P. and Raimondo, M. (1996b). On global performance of approximations to smooth curves using gridded data. Unpublished Manuscript.
- [4] Khintchine, A.Ya. (1935). Zur metrischen Kettenbruch-Theorie. *Compositio Mathematica* **3**, 275–285.
- [5] Khintchine, A.Ya. (1963). *Continued Fractions*. Groningen: Noordhoff.
- [6] Klein, F. (1907). *Ausgewählte Kapitel der Zahlentheorie*. Teubner.
- [7] Korostelev, A.P. and Tsybakov, A.B. (1993). *Minimax Theory of Image Reconstruction*. Springer Lecture Notes in Statistics Vol. 82. Berlin: Springer-Verlag.
- [8] Leveque, W.J. (1956). *Topics in Number Theory*. Volume I. Addison-Wesley, Reading, Mass.
- [9] Lévy, P. (1937). *Théorie de l'Addition des Variables Aléatoires*. Paris: Gauthier-Villars.
- [10] Roth, K.F. (1955). Rational approximations to algebraic numbers. *Mathematika* **2**, 1–20. (Corrigendum *ibid* p. 168.)
- [11] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

# $\ell_1$ computation: An interior monologue

**Roger Koenker**

*Department of Economics, University of Illinois, Champaign, USA*

*Abstract:* Some recent developments on the computation of least absolute error estimators are surveyed and a number of extensions to related problems are suggested. A very elementary example is used to illustrate the basic approach of “interior point” algorithms for solving linear programs. And a simple preprocessing approach for  $\ell_1$  type problems is described. These developments, taken together, have the effect of dramatically improving the efficiency of absolute error computations, making them comparable to least squares methods even in massive datasets.

*Key words:* Linear models, regression quantiles,  $\ell_1$  estimation, computation.

AMS subject classification: 62G05, 62J05, 68A20.

## 1 Why square errors?

Gauss (1823), in what can only be admired as an epitome of “proof by intimidation”, defended his decision to minimize sums of *squared* errors in the following terms:

It is by no means self evident how much loss should be assigned to a given observation error. On the contrary, the matter depends in some part on our own judgment. Clearly we cannot set the loss equal to the error itself; for if positive errors were taken as losses, negative errors would have to represent gains. The size of the loss is better represented by a function that is naturally positive. Since the number of such functions is infinite, it would seem that we should choose the simplest function having this property. That function is unarguably the square, and the principle proposed above results from its adoption.

Laplace has also considered the problem in a similar manner, but he adopted the absolute value of the error as his measure of loss. Now if I am not mistaken, this convention is no less arbitrary than mine. Should an error of double size be considered as tolerable as a single error twice repeated, or worse? Is it better to assign only twice as much influence to a double error or more? The answers are not self-evident, and the problem cannot be resolved by mathematical proofs, but only by an arbitrary decision. Moreover, it cannot be denied that Laplace's convention violates continuity and hence resists analytic treatment, while the results that my contention leads to are distinguished by their wonderful simplicity and generality.

Despite the best efforts of such distinguished advocates as Laplace (1789), Edgeworth (1888), and Kolmogorov (1931), methods of estimation based on minimizing sums of absolute errors have languished in the shade of the edifice that Gauss built on the foundation of least squares. Why? There seem to be at least two elementary reasons. First, the “wonderful simplicity and generality” of squared error has produced an elegant statistical theory of the behavior of least squares estimators which, particularly in its finite-sample form for Gaussian cases, can only inspire awe and envy on the part of advocates of the quantile-esque methods of absolute errors. Some solace may be found in the very critical attack on least-squares based methods by the robustness movement launched by John Tukey in the 1940's. The second, and perhaps even more damaging, is the perception that absolute error estimators are “difficult to compute.” To appreciate that this perception was perfectly valid at the end of the 19th century, one need only read a little of Edgeworth's (1888) own arcane description of his geometric “algorithm” to compute the bivariate least absolute error regression estimator.

With the advent of George Dantzig's simplex algorithm in the late 1940's this situation changed dramatically, and by the mid-50's there were several formulations of the  $\ell_1$  estimator for regression as a linear program and explicit simplex-based programs to compute it. The paper of Wagner (1959) clarified the important role of the  $\ell_1$  dual problem. These efforts culminated in the algorithm of Barrodale and Roberts (1974) which still serves as the  $\ell_1$  algorithm of choice for most statistical computing environments. Contrary to a plethora of dire warnings throughout the literature, about the difficulty of  $\ell_1$  computation this algorithm actually delivers least absolute error regression estimates *faster* than the corresponding least squares algorithms in many packages, including Splus and Stata, for problems of moderate size, up to a few hundred observations. However, for larger problems the Barrodale and Roberts algorithm exhibits  $\mathcal{O}(n^2)$  growth in execution time,

and thus quickly lives up to its slothful reputation. Portnoy (1991) provides a detailed probabilistic complexity analysis for the simplex version of the parametric quantile regression problem which sheds some light on the theoretical rationale for the observed  $\mathcal{O}_p(n^2)$  behavior of the simplex approach. In Portnoy and Koenker (1997), we have shown that combining recent developments on interior point methods for solving linear programs with careful preprocessing can improve both the theoretical and practical performance of  $\ell_1$  regression computations to the point that they are competitive with least squares over the entire range of contemporary problem dimensions.

In this paper I will briefly review these recent developments in  $\ell_1$  computation and then sketch some ideas for extending these developments into a broader range of related problems in statistics.

## 2 Means vs. medians

The most elementary instance of our basic problem may be posed as the simple question: Which is easier to compute, the median or the mean? Surprisingly, the question is deceptively difficult. At the most naive level, it would be immediately recognized that the median has an advantage for computation “by hand”, an attribute implicit in the “median-polish” algorithms suggested by Tukey for robust ANOVA. Somewhat less naively, with modern computers in mind we might contemplate computing the mean in  $\mathcal{O}(n)$  elementary operations ( $n$  additions, and one division), while the median appears to require sorting  $n$  numbers, a task which requires  $\mathcal{O}(n \log n)$  comparisons. Further reflection suggests, however, that the median may not actually require a full sorting of the observations; a cleverly chosen partial sorting may suffice. Considerable further reflection yields the celebrated algorithm of Floyd and Rivest (1975), which manages to compute the median in  $\mathcal{O}(n)$  comparisons. At this point we require a more delicate comparison of the relative effort of additions and comparisons and the precise constants associated with the  $\mathcal{O}(n)$  median algorithm. Since such delicacy seems inherently machine dependent to some degree, we will not attempt to pursue it further here, but will simply note that it is not implausible that a *sophisticated* algorithm for the median could, for  $n$  sufficiently large, outperform the computation of the mean, thus restoring the superiority of the median.

## 3 Simplex for median regression

Portnoy and Koenker (1997) reconsider the problem of solving the  $\ell_1$  re-

gression problem,

$$\min_{b \in \mathbb{R}^p} \sum_{i=1}^n |y_i - x'_i b| \quad (1)$$

which may be formulated as the linear program,

$$\min \{ e'u + e'v \mid y = Xb + u - v, (u, v) \in \mathbb{R}_+^{2n} \}. \quad (2)$$

This problem has the dual formulation

$$\max \{ y'd \mid X'd = 0, \quad d \in [-1, 1]^n \}, \quad (3)$$

or equivalently, setting  $a = d + \frac{1}{2}e$ ,

$$\max \{ y'a \mid X'a = \frac{1}{2}X'e, \quad a \in [0, 1]^n \}. \quad (4)$$

The simplex approach to solving this problem may be briefly described as follows. A  $p$ -element subset of  $\mathcal{N} = \{1, 2, \dots, n\}$  will be denoted by  $h$ , and  $X(h), y(h)$  will denote the submatrix and subvector of  $X, y$  with the corresponding rows and elements identified by  $h$ . Recognizing that solutions to (1) may be characterized as planes which pass through precisely  $p = \dim(b)$  observations, or as convex combinations of such “basic” solutions, we can begin with any such solution, which we may write as,

$$b(h) = X(h)^{-1}y(h). \quad (5)$$

We may regard any such “basic” primal solution as an extreme point of the polyhedral, convex constraint set. A natural algorithmic strategy is then to move to the adjacent vertex of the constraint set in the direction of steepest descent. This transition involves two stages: the first chooses a descent direction by considering the removal of each of the current basic observations and computing the gradient in the resulting direction, then having selected the direction of steepest descent and thus an observation to be removed from the currently active “basic” set, we must find the maximal step length in the chosen direction by searching over the remaining  $n - p$  available observations for a new element to introduce into the “basic” set. Each of these transitions involves an elementary “simplex pivot” matrix operation to update the current basis. The iteration continues in this manner until no direction is found at which point the current  $b(h)$  can be declared optimal.

Sheynin (1973) has noted that Gauss was already aware in 1809 that minimizing absolute errors, as suggested by Boscovich and Laplace, entailed this “zero residual” property. It is therefore tempting to speculate on why

it required another 150 years to develop the “wonderfully simple” idea of moving from vertex to vertex in the direction of steepest descent. One possible explanation for this involves the distinction made by Gill, Murray and Wright (1991) between iterative and direct algorithms. As they put it,

...we consider as *direct* a computation procedure that produces one and only one estimate of the solution, without needing to perform *a posteriori* tests to verify that the solution has been found...In contrast, an iterative method generates a sequence of trial estimates of the solution, called *iterates*. An iterative method includes several elements: an initial estimate of the solution; computable tests to verify whether or not an alleged solution is correct; and a procedure for generating the next iterate in the sequence if the current estimate fails the test.

Thus, the iterative nature of the simplex algorithm makes it rather like a voyage of exploration of the 15th century, sailing into the Atlantic believing that the world was flat, not knowing when, or even if, the voyage would end. Gaussian elimination, on the other hand, made least squares like a trip along a familiar road; at each step one knew exactly how much further effort was necessary. With the emergence of computers in the 1940's, the risk, or uncertainty, of the iterative approach was transferable to the machine, and the spirit of adventure blossomed as investigators put down their pencils and watched the tapes whirl and the lights flicker.

Like the advantage of the median over the mean for hand computations, the simplex algorithm for median regression performs extremely well relative to least squares in problems of modest size. In Figure 1 we can compare performance of the Barrodale and Roberts (1974) algorithm for median regression with the standard least squares algorithm as implemented in the function `lm(y ~ x)` in Splus. For  $p = 4$  and  $n < 2000$ , median regression *à la* Barrodale and Roberts is actually faster than the corresponding least squares computation. As the dimension of the parameter increases, the advantage of  $\ell_1$  over  $L_2$  is somewhat attenuated, but even with  $p = 16$ , there is still an advantage up to sample sizes of about 400.

In larger problems simplex-based computations for median regression pale in comparison with speeds achievable by least squares. In Figure 2 I illustrate this comparison over problems in the range  $20,000 \leq n \leq 120,000$ , and we see that the time required for the modified simplex approach embodied in the Barrodale and Roberts algorithm tends to grow quadratically in  $n$  while the QR factorization approach of `lm` grows only linearly in  $n$ . By sample size  $n = 120,000$  this results in computations of nearly one hour for median regression a procedure which can be carried out in 10-20 seconds by least squares. Is this differential inherent in the linear programming



formulation of the  $\ell_1$  problem, confirming Gauss's claims, or is it simplex that is at fault?

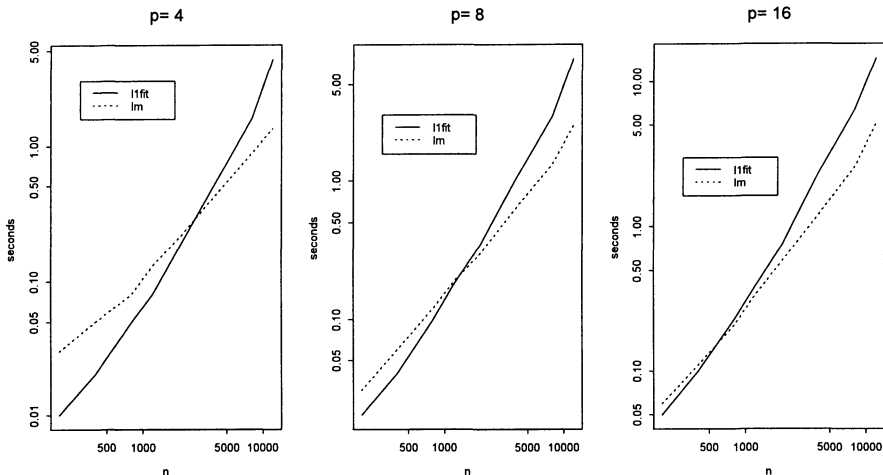


Figure 1: Timing comparison of  $\ell_1$  and  $\ell_2$  algorithms: Times are in seconds for the median of five replications for iid Gaussian data. The parametric dimension of the models is  $p + 1$  with  $p$  indicated above each plot,  $p$  columns are generated randomly and an intercept parameter is appended to the resulting design. Timings were made at 8 design points in  $n$ : 200, 400, 800, 1200, 2000, 4000, 8000, 12000. The solid line represents the results for the simplex-based Barrodale and Roberts algorithm, `l1fit(x,y)` in Splus, and the dotted line represents least squares timings based on `lm(y ~ x)`.

Ironically, one of the great research challenges of numerical analysis of recent decades has been, “Why is simplex so quick?” Examples, notably that of Klee and Minty (1972), have shown that in problems of dimension,  $n$ , simplex can take as many as  $2^n$  simplex pivots, each requiring  $\mathcal{O}(n)$  effort. From this perspective  $\mathcal{O}_p(n^2)$  effort for randomly generated  $\ell_1$  problems appears to be quite brilliant. The recent paper of Shamir (1993) surveys the extensive literature exploring this gap between theoretical worst-case behavior and practical average-case performance. The discussion of simplex in Gill, Murray and Wright (1991) is especially good on this aspect.

## 4 Newton to the max: An elementary example

To illustrate the shortcomings of the simplex method, or indeed of any strategy for solving linear programs which relies on an iterative path along the *exterior* of the constraint set, consider the problem depicted in Figure 3. We have a random polygon whose vertices lie on the unit circle and our objective is to find a point in the polygon that maximizes the sum of its

coordinates, that is, the point furthest north-east in the figure.

Since any point in the polygon can be represented as a convex weighting of the extreme points, the problem may be formulated as

$$\max\{e'u | X'd = u, \ e'd = 1, \ d \in \mathbb{R}_+^n\}, \quad (6)$$

where  $e$  denotes a (conformable) vector of ones,  $X$  is an  $n \times 2$  matrix with rows representing the  $n$  vertices of the polygon and  $d$  is the vector of convex weights to be determined. Eliminating  $u$  we may rewrite (6) somewhat more simply as

$$\max\{s'd | e'd = 1, \ d \in \mathbb{R}_+^n\}, \quad (7)$$

where  $s = Xe$ . This is an extremely simple linear program which serves as a convenient geometric laboratory animal for studying various approaches to solving such problems. Simplex is particularly simple in this context, because the constraint set *is* literally a simplex. If we begin at a random vertex, and move around the polygon until optimality is achieved, we pass through  $O(n)$  vertices in the process. Of course, a random initial vertex is rather naive, and one could do much better with an intelligent “Phase 1” approach that found a *good* initial vertex. In effect we can think of the “interior point” approach we will now describe as a class of methods to accomplish this, rendering unnecessary further travel around the outside of the polygon.

Although prior work in the Soviet literature offered theoretical support for the idea that linear programs could be solved in polynomial time, thus avoiding the pathological exponential growth of the Klee-Minty examples, the paper of Karmarkar (1984) constituted a watershed in the numerical analysis of linear programming. It offered not only a cogent argument for the polynomiality of interior point methods of solving *LP*’s, but also provided for the first time direct evidence that interior point methods were demonstrably faster than simplex in specific, large, practical problems.

To explore several variants of interior point methods we will use our simple polygonal problem. Further details about more general *LP*’s and applications to  $\ell_1$  regression, and quantile regression more generally, may be found in Portnoy and Koenker (1997). The basic approach we will describe to interior point methods for linear programming is set out in the important survey paper by Lustig, Marsden and Shanno (1994). A more detailed exposition may be found in the new monograph of Wright (1996)

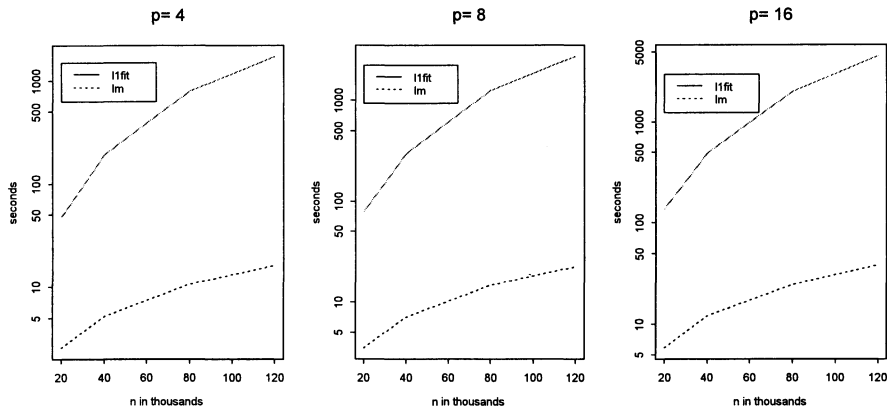


Figure 2: Timing comparison of  $\ell_1$  and  $\ell_2$  algorithms: Times are in seconds for the median of five replications for iid Gaussian data. The parametric dimension of the models is  $p + 1$  with  $p$  indicated above each plot,  $p$  columns are generated randomly and an intercept parameter is appended to the resulting design. Timings were made at 4 design points in  $n$ : 20,000, 40,000, 80,000, 120,000. The solid line represents the results for the simplex-based Barrodale and Roberts algorithm, `l1fit(x,y)` in Splus, and the dotted line represents least squares timings based on `lm(y ~ x)`.

It is an amusing irony, illustrating the spasmodic progress of science, that the most fruitful practical formulation of the interior point revolution of Karmarker (1984) can be traced back to a series of Oslo working papers by Ragnar Frisch in the early 1950's. This work is summarized in Frisch (1956), and was considerably elaborated and extended in the monograph of Fiacco and McCormick (1968). The connection between Karmarker's approach and the earlier literature was developed in Gill, Murray, Saunders, Tomlin and Wright (1986). The basic idea of Frisch was to replace the linear inequality constraints of the  $LP$ , by what may be called a log barrier, or potential, function. Thus, in our example, we may reformulate (7) as,

$$\max\{s'd + \mu \sum_{i=1}^n \log d_i \mid e'd = 1\} \quad (8)$$

where now the barrier term  $\mu \sum \log d_i$  serves as a penalty which keeps us away from the boundary of the positive orthant. By judicious choice of a sequence  $\mu \rightarrow 0$  we might hope to converge to a solution of the original problem.

The salient virtue of the log barrier formulation is that, unlike the original formulation, it yields a differentiable objective function which is con-

sequently attackable by Newton's method. Restricting attention, for the moment, to the primal log-barrier formulation 8 and defining,

$$B(d, u) = s'd + \mu \sum \log d_i \quad (9)$$

we have  $\nabla B = s + \mu D^{-1}e$  and  $\nabla^2 B = -\mu D^{-2}$  where  $D = \text{diag}(d)$ . Thus, at any initial feasible,  $d$ , we have the associated Newton subproblem

$$\max_p \left\{ (s + \mu D^{-1}e)'p - \frac{\mu}{2} p' D^{-2} p \mid e'p = 0 \right\}.$$

This problem has first order conditions

$$\begin{aligned} s + \mu D^{-1}e - \mu D^{-2}p &= ae \\ e'p &= 0 \end{aligned}$$

and multiplying through by  $e'D^2$ , and using the constraint, we have,

$$e'D^2s + \mu e'De = ae'D^2e.$$

Thus solving for the Lagrange multiplier  $\hat{a}$  we obtain the Newton direction

$$p = \mu^{-1} D^2s + De - \hat{a}e \quad (10)$$

where  $\hat{a} = (e'D^2e)^{-1}(e'D^2s + \mu e'De)$ . Pursuing the iteration  $d \leftarrow d + \lambda p$ , thus defined, *with  $\mu$  fixed* until convergence, yields the central path  $d(\mu)$  which describes a yellow brick road to the solution  $d^*$  of the original problem (6). We must be careful to keep the step lengths  $\lambda$  small enough to maintain the interior feasibility of  $d$ . Note that the initial feasible point  $d = e/n$  represents  $d(\infty)$ .

As emphasized by Gonzaga (1992) and others, this central path is a crucial construct for the interior point approach. Algorithms may be usefully evaluated on the basis of how well they are able to follow this path. Clearly, there is some tradeoff between staying close to the path and moving along the path, thus trying to reduce  $\mu$ , iteration by iteration. Improving upon existing techniques for balancing these objectives is the subject of a vast outpouring of current research. Excellent introductions to the subject are provided in the survey paper of Margaret Wright (1992) and the recent monograph of Stephen Wright (1996).

Thus far, we have considered only the primal version of our simple polygonal problem, but it is also advantageous to consider the primal and dual forms together. The dual of (7) is very simple:

$$\min \{a \mid ea - z = s, \quad z \geq 0\}. \quad (11)$$

The scalar,  $a$ , is the Lagrange multiplier on the equality constraint of the primal introduced above, while  $z$  is a vector of “residuals,” or slack variables in the terminology of linear programming. This formulation of the dual exposes the real triviality of the problem – we are simply looking for the maximal element of the vector  $s = Xe$ . This is a very special case of the linear programming formulation of finding any ordinary quantile. But the latter would require us to split  $z$  into its positive and negative parts, and would also introduce upper bounds on the variables,  $d$ , in the primal problem.

Another way to express the central path, one that nicely illuminates the symmetric roles of the primal and dual formulations of the original problem, is to solve the equations,

$$\begin{aligned} e'd &= 1 \\ ea - z &= s \\ Dz &= \mu e. \end{aligned} \tag{12}$$

That solving these equations is equivalent to solving (8) may be immediately seen by writing the first order conditions for (8) as

$$\begin{aligned} e'd &= 1 \\ ea - \mu D^{-1}e &= s, \end{aligned}$$

and then appending the definition  $z = \mu D^{-1}e$ . The equivalence then follows from the negative definiteness of the Hessian  $\nabla^2 B$ . This formulation is also useful in highlighting a crucial interpretation of the log-barrier penalty parameter,  $\mu$ . For any feasible pair  $(z, d)$  we have

$$s'd = a - z'd,$$

so  $z'd$  is equal to the duality gap, the discrepancy between the primal and dual objective functions at the point  $(z, d)$ . At a solution, we have the complementary slackness condition  $z'd = 0$ , thus implying a duality gap of zero. Multiplying through by  $e'$  in the last equation of (12), we may take  $\mu = z'd/n$  as a direct measure of progress toward a solution.

Applying Newton's method to these equations yields

$$\begin{pmatrix} Z & 0 & D \\ e' & 0 & 0 \\ 0 & e & I \end{pmatrix} \begin{pmatrix} p_d \\ p_a \\ p_z \end{pmatrix} = \begin{pmatrix} \mu e - Dz \\ 0 \\ 0 \end{pmatrix},$$

where we have again presumed initial, feasible choices of  $d$  and  $z$ . Solving for  $p_a$  we have

$$\hat{p}_a = (e'Z^{-1}De)^{-1}e'Z^{-1}(Dz - \mu e)$$

which yields the primal-dual Newton direction:

$$p_d = Z^{-1}(\mu e - Dz - Dep_a) \quad (13)$$

$$p_z = ep_a. \quad (14)$$

It is of obvious interest to compare this primal-dual direction with the purely primal step derived above. In order to do so, however, we need to specify an adjustment mechanism for  $\mu$ .

To this end we will now describe an approach suggested by Mehrotra (1992) that has been widely implemented by developers of interior point algorithms, including the interior point algorithm for quantile regression described in Portnoy and Koenker (1997). Given an initial feasible triple  $(d, a, z)$ , consider the affine-scaling Newton direction obtained by evaluating (13) at  $\mu = 0$ . Now compute the step lengths for the primal and dual variables respectively using

$$\lambda_d = \operatorname{argmax}\{\lambda \in [0, 1] | d + \lambda p_d \geq 0\}$$

$$\lambda_z = \operatorname{argmax}\{\lambda \in [0, 1] | z + \lambda p_z \geq 0\}.$$

But rather than precipitously taking this step, Mehrotra suggests adapting the direction somewhat to account for both the “recentering effect” introduced by the  $\mu e$  term in (13) and also for the nonlinearity introduced by the last of the first order conditions.

Consider first the recentering effect. If we contemplate taking a full step in the affine scaling direction we would have,

$$\hat{\mu} = (d + \lambda_d p_d)'(z + \lambda_z p_z)/n,$$

while at the current point we have,

$$\mu = d'z/n.$$

Now, if  $\hat{\mu}$  is considerably smaller than  $\mu$ , it means that the affine scaling direction has brought us considerably closer to the optimality condition of complementary slackness:  $z'd = 0$ . This suggests that the affine scaling direction is favorable, that we should reduce  $\mu$ , in effect downplaying the contribution of the recentering term in the gradient. If, on the other hand,  $\hat{\mu}$  isn't much different than  $\mu$ , it suggests that the affine-scaling direction is unfavorable and that we should leave  $\mu$  alone, taking a step which attempts to bring us back closer to the central path. Repeated Newton steps with  $\mu$  fixed put us exactly on this path. These heuristics are embodied in Mehrotra's proposal to update  $\mu$  by

$$\mu \leftarrow \mu(\hat{\mu}/\mu)^3.$$

To deal with the nonlinearity, Mehrotra (1992) proposed the following “predictor-corrector” approach. A full affine scaling step would entail

$$(d + p_d)'(z + p_z) = d'z + d'p_z + p_d'z + p_d'p_z.$$

The linearization implicit in the Newton step ignores the last term, in effect predicting that since it is of  $\mathcal{O}(\mu^2)$  it can be neglected. But since we have already computed a preliminary direction, we might as well reintroduce this term to correct for the nonlinearity as well to accomplish the recentering. Thus, we compute the modified direction by solving

$$\begin{pmatrix} Z & 0 & D \\ e' & 0 & 0 \\ 0 & e & I \end{pmatrix} \begin{pmatrix} \delta_d \\ \delta_a \\ \delta_z \end{pmatrix} = \begin{pmatrix} \mu e - Dz - P_d p_z \\ 0 \\ 0 \end{pmatrix},$$

where  $P_d = \text{diag}(p_d)$ . This modified Newton direction is then subjected to the same step-length computation and a step is finally taken. It is important in more realistic problem settings that the linear algebra required to compute the solution to the modified step has already been done for the affine scaling step. This usually entails a Cholesky factorization of a matrix which happens to be scalar here, so the modified step can be computed by simply backsolving the same system of linear equations already factored to compute the affine scaling step.

In Figure 3 we provide an example intended to illustrate the advantage of the Mehrotra modified step. The solid line indicates the central path. Starting from the same initial point  $d = e/n$ , the dotted line represents the first affine scaling step. It is successful in the limited sense that it stays very close to the central path, but it only takes a short step toward our final destination. In contrast, the first modified step, indicated by the dashed line, takes us much further. By anticipating the curvature of the central path, it takes a step more than twice the length of the unmodified, affine-scaling step. On the second step the initial affine-scaling step is almost on target, but again somewhat short of the mark. The modified step is more accurately pointed at the desired vertex and is thus, again, able to take a longer step.

It is difficult in a single example like this to convey a sense of the overall performance of these methods. After viewing a large number of realizations of these examples myself, I come away convinced that the Mehrotra modified step consistently improves upon the affine scaling step, a finding that is completely consistent with the theory.

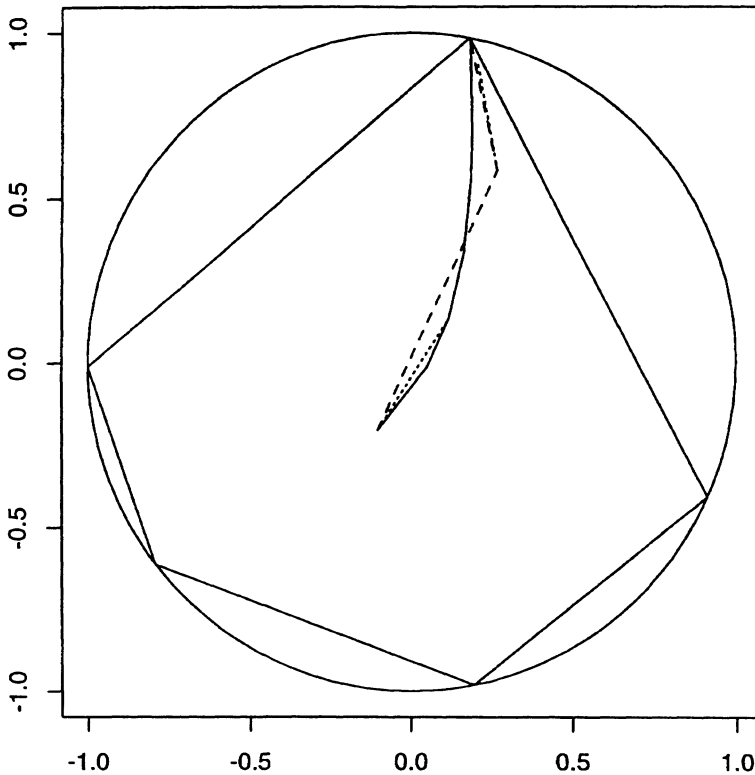


Figure 3: A simple example of interior point methods for linear programming: The figure illustrates a random pentagon of which we would like to find the most northeast vertex. The central path beginning with an equal weighting of the 5 extreme points of the polygon is shown as the solid curved line. The dotted line emanating from this center is the first affine scaling step. The dashed line is the modified Newton direction computed according to the proposal of Mehrotra. Subsequent iterations are unfortunately obscured by the scale of the figure.

In Portnoy and Koenker (1997), we noted that recent work on the probabilistic analysis of the computational complexity of interior point methods suggests that algorithms with  $\mathcal{O}_p(np^3 \log^2 n)$  operations are possible for quantile regression with  $n$  observations and  $p$  parameters. While such performance is considerably better, in large problems, than the observed  $\mathcal{O}_p(n^2 p^2)$  performance of simplex, it is still inferior to the  $\mathcal{O}(np^2)$  complexity of least squares. In the next section I very briefly describe a preprocessing strategy for quantile regression problems that has been successful in further narrowing this computational gap.



## 5 Preprocessing for quantile regression

The idea of preprocessing quantile regression problems described in Portnoy and Koenker (1997) actually preceded the implementation of the interior point methods discussed above. Preprocessing rests on an extremely simple idea: if, by preliminary estimation, or some other form of statistical necromancy, we could determine the signs of a significant group of observations, we could then combine observations with positive residuals into a single “globbed” observation, and similarly glob together the negative observations, so that the original problem,

$$\min \sum_{i=1}^n \rho_{\tau}(y_i - x_i' b) \quad (15)$$

with  $\rho_{\tau}(u) = u(\tau - I(u < 0))$  would be equivalent to,

$$\min \sum_{i \in N \setminus J_L \cup J_H}^n \rho_{\tau}(y_i - x_i' b) + \rho_{\tau}(y_L - x_L' b) + \rho_{\tau}(y_H - x_H' b) \quad (16)$$

where  $N = \{1, 2, \dots, n\}$ ,  $x_K = \sum_{i \in J_K} x_i$  for  $K \in \{K, L\}$  and  $y_L$  and  $y_H$  can be chosen arbitrarily small and large respectively, to ensure that the corresponding residuals on the globbed observations remain negative and positive. In this process we have reduced the problem of  $n$  original observations to  $n - \# \{J_L, J_H\} + 2$  observations so if the cardinality of the  $J$ -sets is large we have gained substantially. Under plausible sampling assumptions we can, based on a preliminary subsample of  $m$  observations, make a prediction region for  $\{x_i \beta : i = 1, 2, \dots, n\}$  of width  $\mathcal{O}(p/\sqrt{m})$ , so assigning observations above this region to  $J_H$  and observations below this region to  $J_L$ , we would have  $M = \mathcal{O}_p(np/\sqrt{m})$  observations falling inside the region. This is illustrated in Figure 4.

Minimizing the computational effort required to compute the preliminary fit based on  $m$  observations plus the effort required for the solution of the globbed problem (16) with  $M$  observations, we obtain  $m^* = \mathcal{O}((np)^{2/3})$ , which under our conjectured performance of the underlying interior point algorithm yields a complexity for the full problem of

$$C = \mathcal{O}_p(n^{2/3} p^3 \log^2 n) + \mathcal{O}(np^2), \quad (17)$$

where the first term comes from the solution of the two quantile regression problems and the second term arises from the computation of the confidence band.

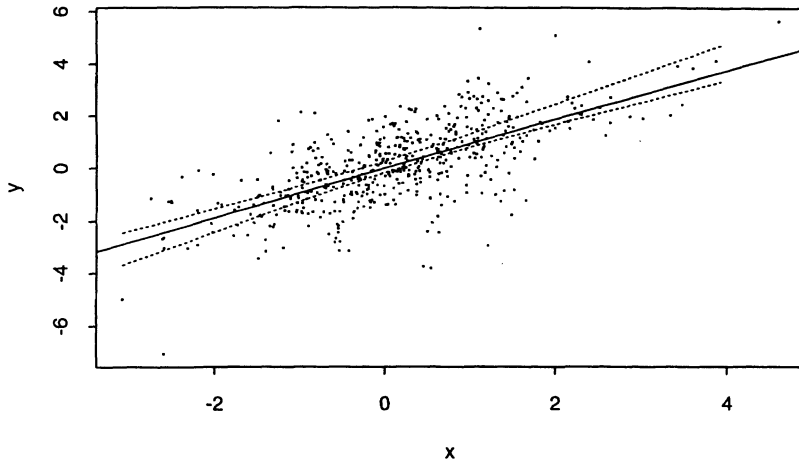


Figure 4: A bivariate example of quantile regression preprocessing: The figure illustrates a bivariate scatter plot of 500 observations with  $y$  conditionally student  $t$  on 10 degrees of freedom. The curved dotted lines describe a confidence band for the response variable based on the median regression fit for a sub-sample of 126 observations. After globbing there are only 107 observations, including the two globbed observations. All the points outside the band are collapsed into this pair of pseudo-observations. The fit to the globbed sample is indicated by the solid line; since it falls inside the band we are assured that the globs are correct and that this solution is identical to a fit of the entire original sample.

Further details are provided in Portnoy and Koenker (1997) and I will comment only briefly here on the important fact that any implementation of this preprocessing approach must verify that the solution to the globbed problem actually vindicates the predicted signs based on the confidence region. Since the simultaneous confidence region can be chosen to assure this with arbitrarily high probability, the eventuality that we may need to repeat the cycle to remedy some inaccurately predicted signs introduces another multiplicative factor which does not affect the orders in probability in the complexity computation.

The crucial consequence of the formal complexity theory and the extensive concomitant empirical testing of our implementation of the algorithm is that the computational effort required for quantile regression can be made comparable with the effort required for least squares over the full range of currently plausible problem dimensions. In the final empirical example of Portnoy and Koenker (1997), we compare timings for a typical large econometric application of quantile regression with  $n = 113,547$  and  $p = 6$ . With the new algorithm, quantile regression estimates take about 10 seconds on a Sparc-Ultra, comparable to the least squares time of 8 seconds. Simplex solution of the same quantile regression problems requires approximately an hour on the same machine.

## 6 Prospects

There are many open questions posed by the rapid development of computational methods for quantile regression. I would like to touch on three topics in this brief final section. The first is applications to inference and the general problem of parametric programming viewed in the light of interior point methods. The second is applications to nonlinear quantile regression. And the third concerns nonparametric applications of quantile regression.

As I have tried to emphasize elsewhere, an important virtue of the simplex approach to  $\ell_1$ -type computation is the direct transition to parametric programming, or sensitivity analysis. Having obtained a solution at one quantile we immediately compute an interval of optimality for this solution, at the endpoints the solution alters. We may then make a simplex pivot which takes us to an adjacent vertex of the constraint set; continuing this process traces out the entire path of solutions to the problem (15) for  $\tau \in [0, 1]$ . Efficient computation of the quantile regression *process* is crucial for the smooth L-statistics described in Koenker and Portnoy (1990), and the corresponding dual process is central to the elegant theory of rank statistics introduced by Gutenbrunner and Jureckova (1992). Very similar computations are required to compute the penalized quantile regression spline estimators introduced in Koenker, Ng and Portnoy (1995) where the degree of smoothing (bandwidth) parameter  $\lambda$  plays the role of  $\tau$ .

The homotopy methods of interior point algorithms also lend themselves naturally to parametric programming. In large problem it may be sufficient to compute solutions on some grid in  $\tau$  or  $\lambda$  and we may thus avoid passing through all the intermediate vertices by tunnelling through the interior of the constraint set, passing directly from one grid point to the next. Algorithms to do this are conceptually straightforward, given the existing research, see for example Monteiro and Mehrotra (1995), but they require some careful engineering.

Non-linear quantile regression, that is quantile regression estimation like (15) with a nonlinear response function replacing the linear predictor  $x_i'\beta$ , are increasingly common in applications. Here too, interior point methods and the preprocessing approaches described above can play a useful role. Some ideas along this line have been already described in Koenker and Park (1996). There is, however, considerable scope for refinement.

Finally, in nonparametric applications of quantile regression there are a wide array of competing methods, all of which can profit from more efficient computational methods for large data sets. This is particularly true of the quantile smoothing spline approach of Koenker, Ng and Portnoy (1995), which offers new challenges in terms of exploiting sparsity in the interior

point matrix computations. This is a topic which has received intense scrutiny in the interior point literature, and there are a number of very promising approaches already available.

We are, I believe, on the verge of overthrowing the long-standing computational disability of  $\ell_1$  methods. In the next century, we may hope that the young statistician looking for improved robustness, or simply for a more complete view of her data, may say of quantile regression, echoing Molly Bloom, "...yes I said yes I will Yes."

## References

- [1] Barrodale, I. and F.D.K. Roberts (1974). Solution of an overdetermined system of equations in the  $\ell_1$  norm. *Commun. ACM* **17**, 319-320.
- [2] Edgeworth, F.Y. (1888). On a new method of reducing observations relating to several quantities. *Philosophical Magazine* **25**, 184-191.
- [3] Fiacco, A.V. and G.P. McCormick (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. New York: Wiley.
- [4] Floyd, R.W. and R.L. Rivest (1975). Expected Time Bounds for Selection. *Commun. ACM* **18**, 165-173.
- [5] Frisch, R. (1956). La Résolution des problèmes de programme linéaire par la méthode du potentiel logarithmique. *Cahiers du Séminaire d'Econometrie*, **4**, 7-20.
- [6] Gauss, C.F. (1823). *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae: Pars Prior*. Translated by G.W. Stewart (1995) as *Theory of the Combination of Observations Least Subject to Error*. Philadelphia: SIAM.
- [7] Gill, P., W. Murray, and M. Wright (1991). *Numerical Linear Algebra and Optimization*. Redwood City: Addison-Wesley.
- [8] Gill, P., W. Murray, M. Saunders, T. Tomlin, and M. Wright (1986). On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method. *Mathematical Programming* **36**, 183-209.
- [9] Gonzaga, C.C. (1992). Path-following methods for linear programming. *SIAM Rev.* **34**, 167-224.
- [10] Gutenbrunner, C. and J. Jureckova (1992). Regression quantile and regression rank score process in the linear model and derived statistics. *Ann. Statist.* **20**, 305-330.
- [11] Karmarkar, N. (1984). A new polynomial time algorithm for linear programming. *Combinatorica* **4**, 373-395.
- [12] Klee, V. and G.J. Minty. How good is the simplex algorithm? In *Inequalities*, Ed. O. Shisha. New York: Academic Press.

- [13] Koenker, R. and G. Bassett (1978). Regression quantiles, *Econometrica*, 46, 33-50.
- [14] Koenker, R. and V. d'Orey (1987,1993). Computing regression quantiles, *Applied Statistics*, **36**, 383-393, and **43**, 410-414.
- [15] Koenker R., P. Ng and S. Portnoy (1995). Quantile Smoothing Splines. *Biometrika* **81**, 673-80.
- [16] Koenker R., and S. Portnoy (1990). L-Estimation for the linear model. *J. Am. Statist. Assoc.* **82**, 851-857.
- [17] Koenker R. and B.J. Park (1996). An interior point algorithm for non-linear quantile regression. *Journal of Econometrics* **71**, 265-283.
- [18] Kolmogorov, A.S. (1931). The method of the median in the theory of errors. *Mat. Sbornik* **38**, 47-50. Reprinted in english in *Selected Works of A. N. Komogorov*, Ed. A.N. Shirayev. Dordrecht: Kluwer.
- [19] Laplace, P.-S. (1789). Sur quelques points du système du monde. Mémoires de l'Académie des Sciences de Paris. Reprinted in *Œuvres Complètes*, Vol. 11. Paris: Gauthier-Villars.
- [20] Lustig, I.J., R.E. Marsden and D.F. Shanno (1992). On umplementing Mehrotra's predictor-corrector interior-point method for linear programming. *SIAM J. Optim.* **2**, 435-449.
- [21] Lustig, I.J., R.E. Marsden and D.F. Shanno (1994). Interior point methods for linear programming: computational state of the art, with discussion. *ORSA Journal on Computing* **6**, 1-36.
- [22] Monteiro, R.D.C., and S. Mehrotra, (1995) A general parametric analysis approach and its implications to sensitivity analysis in interior point methods. Preprint.
- [23] Mehrotra, S. (1992). On the implementation of a primal-dual interior point method. *SIAM J. Optim.* **2**, 575-601.
- [24] Portnoy, S. (1991). Asymptotic behavior of the number of regression quantile breakpoints. *SIAM J. Scient. Statist. Comput.* **12**, 867-883.
- [25] Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: Computation of Squared-error vs. Absolute-error Estimators. *Statistical Science*. Forthcoming.
- [26] Shamir, R. (1993). Probabilistic analysis in linear programming. *Statistical Science* **8**, 57-64.
- [27] Wagner, H.M., (1959). Linear programming techniques for regression analysis. *J. Am. Statist. Assoc.* **54**, 206-212.
- [28] Wright, M.H. (1992). Interior methods for constrained optimization. *Acta Numerica*, 341-407.
- [29] Wright, S.J. (1996). *Primal-Dual Interior-Point Methods*. Philadelphia: SIAM.

# Regression rank statistics

C. Gutenbrunner

*University of Marburg, Germany*

*Abstract:* This article deals with a family of implicitly defined rank statistics, which are designed to make inference on general linear hypotheses in a large class of nonparametric extensions of the classical linear model. The new rank statistics are defined via the solutions of a continuous family of minimization problems. For simple designs, the procedure leads to the classical rank statistics.

*Key words:* Rank regression, logistic regression, robust estimation, transformation models, linear models.

AMS subject classification: Primary 62G05, 62G10, 62G20; secondary 62G30, 62J02.

## 1 Introduction

For a given known c.d.f.  $F_0$  with continuous positive density  $f_0$  and finite second moment, let us first consider the classical parametric linear model

$$M^{\text{Par}}(F_0) : Y_i \sim F_0(t - \mu_i), \quad \mu_i = \boldsymbol{\beta}' \mathbf{x}_i \quad (1)$$

where  $Y_i$ ,  $1 \leq i \leq n$  are independent responses and the vectors  $\mathbf{x}_i$  represent design conditions and covariables (we assume that the first component  $x_{i1}$  of the  $\mathbf{x}_i$  is 1 corresponding to the intercept and denote with  $\mathbf{X} = (\mathbf{x}'_1, \dots, \mathbf{x}'_n)'$  the design matrix). Usually, in such models one is interested in linear hypotheses of the form

$$H_0^{\text{Par}}(F_0) : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}. \quad (2)$$

It is well known, that this model is not *invariant* w.r.t. nonlinear increasing transformations of the response, that is, if  $m(t)$  is a nonlinear increasing function, then the transformed responses  $m(Y_i)$  in general do not follow

a linear model of the form described above (in the following, we use the terms *invariant* or *ordinal invariant* as short terms for *invariant against strictly increasing transformations*). On the other hand, in practice it often is obvious, that the observed responses do not come from a linear model (1), but there might be an unknown nonlinear monotone transformation, such that the transformed observations do.

The smallest invariant model containing  $M^{\text{Par}}(F_0)$  is the semiparametric transformation model

$$M^{\text{SPar}}(F_0) : Y_i \sim F_0(a(t) - \mu_i), \quad \mu_i = \beta' \mathbf{x}_i, \quad (3)$$

where the unknown increasing function  $a(t)$  introduces a nonparametric component.  $\mu = (\mu_1, \dots, \mu_n)'$  is still identifiable up to an intercept, i.e. a constant multiple of  $\mathbf{1}_n$ . The null hypothesis (2) in the context of model  $M^{\text{SPar}}(F_0)$  will be denoted with  $H_0^{\text{SPar}}(F_0)$ .

Rank statistics for inference within this kind of model were studied for example in Pettitt (1982, 1983, 1987), Doksum (1987), Cuzick (1988), Tsukahara (1992), see also Bickel et al. (1993) and Chauduri et al. (1994). They are based on different approximations of Hoeffding's formula (Hoeffding, 1951) for the partial or marginal likelihood of the ranks,

$$n! \mathbb{P}_\beta(\mathbf{R}(\mathbf{Y}) = \mathbf{r}) = \mathbb{E} \left( \prod_{i=1}^n \frac{f_0(F_0^{-1}(U_{n:r_i}) - \beta' \mathbf{x}_i)}{f_0(F_0^{-1}(U_{n:r_i}))} \right) \quad (4)$$

in which  $U_{n:l}$  is the  $l$ -th order statistic in a sample of  $n$  i.i.d. variables  $U_{n1}, \dots, U_{nn}$  distributed uniformly on  $(0, 1)$ , and  $\mathbf{r} = (r_1, \dots, r_n)'$ .

Note, that in (4) only  $\beta$  is unknown, that is, reduction to the maximally invariant vector of ranks leads to a parametric likelihood, which however in general is numerically difficult. One prominent counter-example is the *proportional hazards model* ( $F_0$  the extreme value c.d.f.), in which (4) becomes Cox's (1972) partial likelihood (for the case without censoring or ties)

$$n! \mathbb{P}_\beta(\mathbf{R}(\mathbf{Y}) = \mathbf{r}) = \prod_{i=1}^n \frac{\exp(\beta' \mathbf{x}_{(i)})}{\sum_{j=i}^n \exp(\beta' \mathbf{x}_{(j)})},$$

where  $\mathbf{x}_{(i)}$  corresponds to the  $i$ -th largest response.

Pettitt treated approximations to (4) for  $F_0$  normal, whereas Doksum and Cuzick used different approximations for the general case, which however still are numerical quite difficult. For the  $F_0 = \Phi$  (normal) case, which is one of the cases with an explicit expression for the expectation of the ranks, Pettitt also proposed an alternative to maximizing a likelihood, namely to minimize

$$\sum_i [R_i - \mathbb{E}_\beta(R_i)]^2, \quad \text{with} \quad \mathbb{E}_\beta(R_i) = 1 + \sum_{j \neq i} \Phi[(\mathbf{x}_i - \mathbf{x}_j)' \beta / \sqrt{2}].$$

## 2 Statistical model

In the present paper, we shall generalize (3) to a nonparametric model  $M^{\text{NPar}}(F_0)$ , which may be regarded as the synthesis of (3) with a nonparametric model proposed by Akritas and Arnold (1994) in the case of  $\mathbf{X}$  representing a factorial design. The extension compared to (4) consists in replacing inside  $F_0$  the functions  $\mu_i(t) = a(t) - \boldsymbol{\beta}' \mathbf{x}_i$  with the more general functions  $\mu_i(t) = \mathbf{x}_i' \boldsymbol{\beta}(t)$  where  $\boldsymbol{\beta}(t)$  is any smooth function of  $t$ , such that the  $\mu_i(t)$  are strictly increasing:

$$M^{\text{NPar}}(F_0) : Y_i \sim F_0(\mathbf{x}_i' \boldsymbol{\beta}(t)). \quad (5)$$

The semiparametric transformation model (3) is contained as special case  $\boldsymbol{\beta}(t) = a(t)\mathbf{e}_1 - \boldsymbol{\beta}$ , the Akritas-Arnold (1994) model is obtained from (5) by choosing  $F_0(t) = t$ , the c.d.f. of the uniform distribution. However this latter choice does not satisfy our requirement that the density  $f_0$  should be positive on the whole real line. This causes the compatibility conditions (8)-(11) to fail for the Akritas-Arnold model.

Introducing the *link function*  $h = F_0^{-1}$ , and denoting with  $F_i$  the c.d.f. of  $Y_i$ ,  $h \circ \mathbf{F} = (h \circ F_1, \dots, h \circ F_n)'$ , we may state (5) also in the form that  $h \circ \mathbf{F}$  must be continuously differentiable and satisfy for all  $t \in \mathbb{R}$  a linear condition

$$h \circ \mathbf{F}(t) \in \mathbf{L} = \mathbf{X}\mathbb{R}^p = \{\mathbf{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}.$$

A natural extension of the linear hypothesis (2) is given by

$$H_0^{\text{NPar}}(F_0) : \mathbf{C}\boldsymbol{\beta}(t) = \mathbf{0} \quad \text{for all } t, \quad \text{or equivalently} \quad (6)$$

$$H_0^{\text{NPar}}(F_0) : h \circ \mathbf{F}(t) \in \mathbf{L}_0 = \{\mathbf{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p, \mathbf{C}\boldsymbol{\beta} = \mathbf{0}\} \quad \text{for all } t. \quad (7)$$

These definitions imply the *compatibility properties*

$$M^{\text{Par}}(F_0) \subset M^{\text{SPar}}(F_0) \subset M^{\text{NPar}}(F_0), \quad (8)$$

$$H_0^{\text{SPar}}(F_0) = M^{\text{SPar}}(F_0) \cap H_0^{\text{NPar}}(F_0), \quad (9)$$

$$H_0^{\text{Par}}(F_0) = M^{\text{Par}}(F_0) \cap H_0^{\text{SPar}}(F_0) \quad (10)$$

$$= M^{\text{Par}}(F_0) \cap H_0^{\text{NPar}}(F_0) \quad (11)$$

The models presented above should not be confused with the *semiparametric shift model*  $M^{\text{SShft}} = \bigcup_{F_0 \in \mathcal{F}} M^{\text{Par}}(F_0)$ ,  $\mathcal{F}$  a set of real c.d.f., i.e.

$$M^{\text{SShft}} : Y_i \sim F(t - \mu_i), \quad F \in \mathcal{F} \text{ unknown.} \quad (12)$$

This model is not invariant in our sense (w.r.t. monotone transformations). It is closely related to a large class of rank tests, namely the procedures



collected under the name “ranking after alignment” (overviews in Adichie (1984), Puri and Sen (1985), further references see there).

For example, in the  $k$ -sample problem,  $M^{\text{NPar}}(F_0)$  is larger than  $M^{\text{SShft}}$  (requiring only groupwise identical distribution but not the assumption that the between-group differences just shift the c.d.f.), whereas the hypothesis

$$H_0^{\text{NPar}}(F_0) : \beta_2(t) = \dots = \beta_p(t) = 0 \text{ for all } t$$

is identical with the classical hypothesis of no group differences.

Besides the property of being an invariant extension of the parametric linear model,  $M^{\text{SPar}}(F_0)$  and  $M^{\text{NPar}}(F_0)$  have another interesting feature: They are characterized by the property, that any discretization of the response  $Y$  leads to the class of *discrete ordinal regression models* commonly named “cumulative logit” models (but the link function need not be the logit link, it is just  $h$ ).  $M^{\text{SPar}}(F_0)$  corresponds to the model with the “parallel regression lines” assumption (which is tested e.g. by SAS procedure LOGISTIC). This model was considered e.g. by McCullagh (1980) and Anderson and Philips (1981).  $M^{\text{NPar}}(F_0)$  corresponds to the model without parallel regression lines assumption, which was considered e.g. by Williams and Grizzle (1972).

**Example 1** *Let us consider the  $2 \times 2$  factorial design with a continuous covariate  $x$ . The nonparametric transformation model (5) in this case is*

$$\mathbb{P}(Y_{ijk} \leq t) = F_0(\alpha(t) + \gamma_{1i}(t) + \gamma_{2j}(t) + \gamma_{3ij}(t) + \gamma_4(t)x_{ijk}),$$

$\alpha, \gamma_{1i}, \gamma_{2j}, \gamma_{3ij}, \gamma_4$  smooth functions of  $t$ , and

$$H_0 : \gamma_{3ij} \equiv 0 \quad \text{or} \quad H_0 : \gamma_4 \equiv 0.$$

*Within the semiparametric transformation model, only  $\alpha$  may depend on  $t$ , within the parametric submodel  $\alpha(t)$  must be a linear function of  $t$ . The Akritas-Arnold-model corresponds to  $F_0(t) = t$  on  $(0, 1)$ .*

### 3 Regression rank statistics

We define the *regression rank score*  $\hat{\beta}(Y_i)$  of response  $Y_i$  by the nonlinear regression equation

$$\sum_{j=1}^n F_0(\mathbf{x}_j' \hat{\beta}(Y_i)) \mathbf{x}_j = \sum_{j=1}^n c(Y_i - Y_j) \mathbf{x}_j, \quad (13)$$

where  $c(t) = 0.5I[t = 0] + I[t > 0]$ .

**Proposition 1** *If  $f_0$  is positive on  $\mathbb{R}$  and has finite expectation, the solution of (13) exists and is unique.*

**Proof:** The solution characterizes the minimum of the strictly convex function of  $\beta$ :

$$G_i(\beta) = \sum_{j=1}^n \left( \int_{h(0)}^{\mathbf{x}'_j \beta} [F_0(t) - c(Y_i - Y_j)] dt \right), \quad (14)$$

which is bounded from below because  $\int |t| dF_0(t) < \infty$ .  $\square$

The solution might be infinite if there exists  $\mu \in \mathbf{L}$  such that  $\text{sign}(\mu_j) = 2c(Y_i - Y_j) - 1$  for all  $j$ , that is, the  $\mathbf{x}_j$  corresponding to  $Y_j \leq Y_i$  are separated from those corresponding to  $Y_j > Y_i$  by some hyperplane  $\{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{x}' \beta^* = 0\}$ . We may avoid infinite solutions by replacing  $F_0$  with  $[(n+1)F_0 - 0.5]/n$  in (13) and (14).

Only in the special case of  $h$  being the logit link, (13) defines the maximum likelihood estimator. In the one sample case ( $Y_i$  i.i.d.,  $\mathbf{x}_i \equiv 1$ ), (13) reduces to

$$nF_0(\hat{\beta}(Y_i)) = \sum_{j=1}^n c(Y_i - Y_j) = R(Y_i), \quad \text{hence}$$

$$\hat{\beta}(Y_i) = h(R(Y_i)/n), \quad R(Y_i) = \text{rank of } Y_i.$$

In the  $k$ -sample case,  $\hat{\beta}(Y_i)$  is the vector of rank scores of  $Y_i$  within the  $k$  samples,

$$\hat{\beta}(Y_i) = \left( h \left[ \frac{R_1(Y_i)}{n_1} \right], \dots, h \left[ \frac{R_k(Y_i)}{n_k} \right] \right)'. \quad (15)$$

In the general case, it might appear at the first glance that computation of all  $n$  regression rank scores could be a time consuming task. This is not true however, if they are computed sequentially in their natural order: If  $Y_{n:1} < \dots < Y_{n:n}$  denotes the ordered sample, one should compute the *ordered regression rank scores*  $\hat{\beta}(Y_{n:i})$  sequentially:  $\hat{\beta}(Y_{n:i})$  is computed from data  $c(Y_{n:i} - Y_j)$ ,  $\hat{\beta}(Y_{n:i+1})$  is computed from data  $c(Y_{n:i+1} - Y_j)$ , which differ from the first set only by adding 0.5 to two components. Consequently, only few iterations will be necessary to compute  $\hat{\beta}(Y_{n:i+1})$  when using  $\hat{\beta}(Y_{n:i})$  as starting value. Hence the actual effort for computation of all solutions is only the effort of computing one nonlinear regression plus  $O(n)$ . The regression rank statistics we define below, do not use the solutions at extreme order statistics. It is convenient to start the iteration process at the sample median and to proceed in both directions, until the solutions at the  $\epsilon \times 100\%$ - and the  $(1 - \epsilon) \times 100\%$ - sample quantiles are obtained.

In general, it is useful for both motivation and investigation of (13), to replace there  $Y_i$  respective  $Y_{n:i}$  with a continuously varying parameter  $t$ :

$$\sum_{j=1}^n F_0(\mathbf{x}_j' \hat{\boldsymbol{\beta}}(t)) \mathbf{x}_j = \sum_{j=1}^n c(t - Y_j) \mathbf{x}_j. \quad (16)$$

The path  $t \mapsto \hat{\boldsymbol{\beta}}(t)$  jumps at the points  $t_i = Y_{n:i} = \hat{H}^{-1}(i/n)$ ,  $u \mapsto \hat{\boldsymbol{\beta}} \circ \hat{H}^{-1}(u)$  jumps at  $u_i = i/n$ , where in the usual notation,  $\hat{H}$  is the empirical c.d.f. of the total (pooled) sample and  $H = n^{-1} \sum_{i=1}^n F_i$ . A key role will be played by the *regression rank score process* for the transformation model,

$$u \mapsto n^{1/2}(\hat{\boldsymbol{\beta}} \circ \hat{H}^{-1}(u) - \boldsymbol{\beta} \circ H^{-1}(u)). \quad (17)$$

**Remark 1** *As it was mentioned already, (13) is the ML-equation only if  $h$  is the logistic link function. A worthwhile alternative to the present approach consists in replacing (13) by the ML-equation for any link function. This could improve efficiency, but at the cost of additional assumptions, like strong unimodality of  $f_0$ , to guarantee uniqueness of the solution. Note, that our assumptions, requiring only positivity, continuity and finite first moment for  $f_0$ , are very weak.*

**Remark 2** *In Gutenbrunner and Jurecková (1992), we defined regression rank scores for the semiparametric shift model  $M^{\text{SShift}}$  in a different way, due to the different nature of the statistical model and the associated invariance requirements (invariance w.r.t.  $p$ -dimensional affine transformations versus invariance w.r.t. componentwise monotone transformations). The common name is used because of the common purpose, namely to define rank statistics for the general linear model.*

We shall define now generalized rank statistics which are appropriate for testing linear hypotheses  $H_0^{\text{NPar}}(F_0)$  within model  $M^{\text{NPar}}(F_0)$ . We call these statistics, which are linear combinations of ordered regression rank scores, shortly *regression rank statistics*.

More specifically, we focus here on weighted averages

$$\hat{\mathbf{T}} = \sum_{i=1}^n w_{ni} \hat{\boldsymbol{\beta}}(Y_{n:i}),$$

with weights satisfying  $\sum_{i=1}^n w_{ni} = 1$ ,  $w_{ni} \geq 0$ ,  $w_{ni} = 0$  if not  $\epsilon \leq i/n \leq 1 - \epsilon$  for some  $\epsilon > 0$ .

We assume here, that the weights are generated by a score function  $J$  via

$$w_{ni} = J\left(\frac{i+1}{n}\right) - J\left(\frac{i}{n}\right),$$

and that  $J$  is the c.d.f. of a probability measure, which is concentrated on a compact subset of the open unit interval.

To stress the dependence of  $\hat{\mathbf{T}}$  on  $J$  and its implicit dependence on the link function  $h$ , we shall also write sometimes  $\hat{\mathbf{T}} = \hat{\mathbf{T}}(h, J)$ . Writing  $\hat{H}$  for the empirical c.d.f. of the total sample,  $Y_{n:i} = \hat{H}^{-1}(i/n)$  and  $H = n^{-1} \sum_{i=1}^n F_i$ , we have the representation

$$\hat{\mathbf{T}}(h, J) = \sum_{i=1}^n w_{ni} \hat{\beta}(Y_{n:i}) = \int \hat{\beta} \circ \hat{H}^{-1}(u) dJ(u) = \int \hat{\beta}(t) dJ \circ \hat{H}(t) \quad (18)$$

corresponding to the functional

$$\mathbf{T}(h, J) = \int \beta \circ H^{-1}(u) dJ(u) = \int \beta(t) dJ \circ H(t). \quad (19)$$

Within  $M^{\text{SPar}}(F_0)$ , (19) reduces to

$$\mathbf{T}(h, J) = \int a(t) dJ \circ H(t) \mathbf{e}_1 - \beta$$

From (19) it is clear, that the null hypothesis (6)/(7) implies  $\mathbf{CT}(h, J) = \mathbf{0}$ .

The two “score functions”  $h$  and  $J$  determine the rank statistic  $\hat{\mathbf{T}}(h, J)$  in an asymmetric way: While  $h$  is crucial for the statistical model and must be specified correctly in order to obtain consistent estimators and tests,  $J$  merely determines efficiency properties like the score function of ordinary linear rank statistics.

The basic idea, to replace one complicated estimating equation with a continuous family of simple equations and to take a weighted average of the family of solutions instead the one solution of the complicated equation, is not new: Koenker and Bassett (1978) extended sample quantiles to linear shift model  $M^{\text{SShft}}$  (12) using an analogous approach, Koenker and Portnoy (1987) and Gutenbrunner and Jurečková (1992) considered linear combinations of solutions.

**Example 2** *Taking the design from Example 1, but without the continuous covariate, the regression ranks take the explicit form (15), since the model is saturated. The components of rank statistic of type  $\hat{\mathbf{T}}$  for testing  $H_0 : \gamma_{3ij} \equiv 0$  hence can be expressed directly as*

$$\hat{T}_{ij}(h, J) = \sum_{l=1}^n w_l h[R_{ij}(Y_{n:l})],$$

where  $R_{ij}(Y_{n:l})$  is the rank of the  $l$ -th largest pooled observation w.r.t. the  $ij$ -th group.

For comparison with other rank statistics it is convenient to sum by parts and express the statistic as a function of the ranks  $R_{ijk} = n\widehat{H}(Y_{ijk})$  within the combined sample. This leads to

$$\begin{aligned}\widehat{T}_{ij}(h, J) &= - \int (J \circ \widehat{H} - J \circ \widehat{F}_{ij}) dh \circ \widehat{F}_{ij} \\ &= - \sum_{k=1}^{n_{ij}} v_{ijk} \left[ J \left( \frac{(n_{ij}/n)R_{ij:k} + 0.5}{n_{ij} + 1} \right) - J \left( \frac{k}{n_{ij} + 1} \right) \right],\end{aligned}$$

where  $R_{ij:k}$  is the  $k$ -th largest rank in group  $ij$  and

$$v_{ijk} = h \left( \frac{k + 0.5}{n_{ij} + 1} \right) - h \left( \frac{k - 0.5}{n_{ij} + 1} \right),$$

using appropriate versions of empirical c.d.f. and ranks.

**Example 3 (median scores):** *The simplest score function in our context is the median score function  $J(u) = I[u \geq 0.5]$ . It is particularly interesting because our results in this case state that the proper generalization of the median test to tests for general linear hypotheses consists in a very pragmatic procedure: dichotomize the continuous response at the pooled median and compute a categorical regression (like logistic or probit regression). Our results in Section 4 on the asymptotics imply that not knowing the true median (e.g. the data dependent dichotomization point) introduces an additional random vector to the estimator of  $\beta(0.5)$ , which is contained in the linear space generated by the null hypothesis and hence does not affect test statistics based on contrasts orthogonal to that space. In the semiparametric transformation model  $M^{\text{SPar}}(F_0)$  the additional random component asymptotically is proportional to the intercept, the slope components of  $\widehat{\beta}$  are not affected. In the nonparametric extension  $M^{\text{NPar}}(F_0)$ , it is proportional to the derivative  $\dot{\beta}(t)$  of  $t \mapsto \beta(t)$  at  $t = H^{-1}(0.5)$ , as follows from (25)- (28).*

**Example 4 (Steam data from Draper and Smith, 1981):** *Doksum (1987) used these data to demonstrate his Monte Carlo approximation for the MPLE (approximately maximizing Hoeffding's partial likelihood (4)). The data are an example of a simple linear regression with a good fit of the parametric linear model, hence a case where the ordinary least squares estimator (LSE) is appropriate.*

The response  $Y$  is pounds of steam per month needed by a power plant, the regressor  $x$  average atmospheric temperature in degree Fahrenheit. In the usual notation

$$Y_i = \alpha + \tilde{\beta}x_i + \sigma\epsilon_i,$$

one has to be aware that our  $\beta' = (\beta_1, \beta_2)$  from the transformation model is  $(\alpha/\sigma, \tilde{\beta}/\sigma)$ , because  $F_0((t - \alpha - x\tilde{\beta})/\sigma) = F_0(t/\sigma - \alpha/\sigma - x\tilde{\beta}/\sigma) = F_0(a(t) - \alpha/\sigma - x\tilde{\beta}/\sigma)$ .

We compare examples of our regression rank statistics with the LSE and Doksum's likelihood sampling estimator. The second, third and fifth regression rank estimators correspond to the score function  $J(u) = (-\frac{1}{2}) \vee 3(u - \frac{1}{2}) \wedge \frac{1}{2}$ , defining a trimmed mean type statistic. There is a remarkable gap between the LSE and our estimators on one hand and Doksum's estimator on the other hand. Also replacing estimating equation (13) by the ML-equation for probit regression takes the estimator nearer to the LSE.

The estimates in descending order:

- 0.092 LSE (95%-confidence limits:  $[-.136, -.048]$ )
- 0.090 trimmed mean of regression rank scores, *probit link function*, using ML-equation for dichotomized data instead of (13) (trimming: 33% from both sides).
- 0.084 trimmed mean of regression rank scores, *logit link function* (trimming: 33% from both sides).
- 0.082 regression rank score median, *probit link function*, using (13).
- 0.081 trimmed mean of regression rank scores, *probit link function*, using (13) (trimming: 33% from both sides)
- 0.064 Doksum's likelihood sampling estimator

## 4 Asymptotic representations

In this section we show that the asymptotic representation of regression rank statistics  $\hat{\mathbf{T}}(h, J)$  (18) shares important properties of the corresponding representations for linear rank statistics as it was recently developed by Akritas and Arnold (1994), Akritas and Brunner (1996) and others.

According to the asymptotic behaviour of the design we assume

$$\|\mathbf{X}\|_{\infty} = o(n^{1/2}) \quad \text{as } n \rightarrow \infty, \quad (20)$$

$$\|(n^{-1}\mathbf{X}'\mathbf{X})^{-1}\|_2 = O(1) \quad \text{as } n \rightarrow \infty, \quad (21)$$

$$\sup_{n>0} n^{-1} \sum_{i=1}^n I[\|\mathbf{x}_i\| > K] \rightarrow 0 \quad \text{as } K \rightarrow \infty. \quad (22)$$

For  $F_i(t) = F_0(\mathbf{x}_i'\beta(t))$  we assume a continuous density  $f_i(t)$ . This assumption is equivalent to continuous differentiability of  $\beta(t)$  w.r.t.  $t$ . We denote the derivative w.r.t  $t$  as  $\dot{\beta}(t)$ .

Much in the spirit of Pyke and Shorak (1968), we shall expand the regression rank process (17) into the difference of two empirical processes.

Writing  $\mathbf{Z}(t)$  for the vector with components  $c(t - Y_i)$ , we may write (16) as

$$\begin{aligned} \mathbf{g}_t(\widehat{\boldsymbol{\beta}}(t)) &= \mathbf{0}, \quad \text{where} \\ \mathbf{g}_t(\boldsymbol{\beta}) &= \mathbf{X}'[F_0 \circ \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}(t)], \quad \boldsymbol{\beta} \in \mathbb{R}^p. \end{aligned}$$

The derivative  $\mathbf{D}_t = (\partial/\partial\boldsymbol{\beta})\mathbf{g}_t(\boldsymbol{\beta}(t))$  is

$$\mathbf{D}_t = \mathbf{X}' \text{diag}[f_0 \circ \mathbf{X}\boldsymbol{\beta}(t)]\mathbf{X}. \quad (23)$$

The following steps are standard calculations: Denoting with  $o^*(\cdot)$  and  $o_p^*(\cdot)$  convergence uniformly in  $t$  for  $t$  in compact sets,

$$\begin{aligned} \mathbf{0} &= \mathbf{g}_t \circ \widehat{\boldsymbol{\beta}}(t) = \mathbf{g}_t \circ \boldsymbol{\beta}(t) + \mathbf{D}_t[\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)] + o^*(\|\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)\|) \\ &= \mathbf{X}'(\mathbf{F}(t) - \mathbf{Z}(t)) + \mathbf{D}_t[\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)] + o^*(\|\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)\|) \end{aligned}$$

hence

$$n^{1/2}(\widehat{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}(t)) = \mathbf{W}_n(t) + o_p^*(1),$$

with the empirical process

$$\mathbf{W}_n(t) = n^{1/2}\mathbf{D}_t^{-1}\mathbf{X}'(\mathbf{Z}(t) - \mathbf{F}(t)). \quad (24)$$

We used here, that  $f_0(\mathbf{x}_i'\boldsymbol{\beta}(t))$  is bounded away from zero for  $t$  in compact sets. The next routine step is

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\beta}} \circ \widehat{H}^{-1} - \boldsymbol{\beta} \circ H^{-1}) &= \mathbf{W}_n \circ H^{-1} + \sqrt{n}(\boldsymbol{\beta} \circ \widehat{H}^{-1} - \boldsymbol{\beta} \circ H^{-1}) + o_p^*(1) \\ &= \mathbf{W}_n \circ H^{-1} + \sqrt{n}\dot{\boldsymbol{\beta}} \circ H^{-1}(\widehat{H}^{-1} - H^{-1}) + o_p^*(1) \\ &= \mathbf{W}_n \circ H^{-1} - \sqrt{n}\frac{\widehat{H} \circ H^{-1} - id}{\dot{H} \circ H^{-1}}\dot{\boldsymbol{\beta}} \circ H^{-1} + o_p^*(1) \\ &= \mathbf{W}_n \circ H^{-1} - \mathbf{V}_n \circ H^{-1} + o_p^*(1) \end{aligned} \quad (25)$$

with the second empirical process

$$\mathbf{V}_n(t) = \sqrt{n}\frac{\widehat{H}(t) - H(t)}{\dot{H}(t)}\dot{\boldsymbol{\beta}}(t). \quad (26)$$

In (25), the asymptotic replacement of  $\sqrt{n}(\widehat{H}^{-1} - H^{-1})$  by  $-\sqrt{n}\frac{(\widehat{H} \circ H^{-1} - id)}{\dot{H} \circ H^{-1}}$  follows the argument given e.g. in Serfling (1980), sect. 2.8.3., p. 112.

The vector  $\dot{H}(t)^{-1}\dot{\boldsymbol{\beta}}(t)$  may also be written as  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{d}(t)$ ,  $d_i(t) = (f_i(t)/\bar{f}(\cdot(t)))h'(F_i(t))$ . If  $H_0^{\text{NPar}}(F_0)$  is true, the derivative  $\mathbf{C}\dot{\boldsymbol{\beta}}(t)$  of  $\mathbf{0} \equiv \mathbf{C}\boldsymbol{\beta}(t)$  is zero, therefore  $\mathbf{V}_n(t)$  has the important property

$$\mathbf{C}\mathbf{V}_n(t) \equiv \mathbf{0} \quad \text{under } H_0^{\text{NPar}}(F_0).$$

Hence, defining

$$\mathbf{M}_i(t) = \int (I[t \leq s] - F_i(s))(\mathbf{D}_s^{-1}\mathbf{x}_i - \dot{H}(s)^{-1}\dot{\beta}(s)) dJ \circ H(s) \quad \text{and}$$

$$\mathbf{M}_i^*(t) = \int (I[t \leq s] - F_i(s))(\mathbf{D}_s^{-1}\mathbf{x}_i) dJ \circ H(s),$$

we have sketched the proof of the asymptotic representation stated in the following

**Theorem 1** *Within model  $M^{\text{NPar}}(F_0)$ , if the density  $f_0$  is continuous, positive on  $\mathbb{R}$  and has finite expectation, the score function  $J$  is trimming and of bounded variation and conditions (20)-(22) are satisfied, then*

$$\hat{\mathbf{T}}(h, J) = \mathbf{T}(h, J) + n^{-1} \sum_{i=1}^n \mathbf{M}_i(Y_i) + o_p(n^{-1/2}), \quad (27)$$

and, under  $H_0^{\text{NPar}}(F_0)$ ,

$$\mathbf{C}\hat{\mathbf{T}}(h, J) = n^{-1} \sum_{i=1}^n \mathbf{C}\mathbf{M}_i^*(Y_i) + o_p(n^{-1/2}). \quad (28)$$

The regression rank score process (17) has the asymptotic representation

$$\sqrt{n}(\hat{\beta} \circ \hat{H}^{-1}(u) - \beta \circ H^{-1}(u)) = \mathbf{W}_n \circ H^{-1}(u) - \mathbf{V}_n \circ H^{-1}(u) + o_p^*(1),$$

where  $o_p^*(1)$  denotes approximation uniformly in  $u$  for  $u$  in compact subsets of  $(0,1)$  and the empirical processes  $\mathbf{W}_n$  and  $\mathbf{V}_n$  are given by (24) and (26).  $\square$

(27) and (28) imply multivariate asymptotic normality with an asymptotic covariance matrix that under  $H_0^{\text{NPar}}(F_0)$  has a simplified structure that follows from the covariance function  $\mathbf{K}(s, t)$  of  $\mathbf{W}_n$ :

$$\mathbf{K}(s, t) = \mathbf{D}_s^{-1} \Delta(s, t) \mathbf{D}_t^{-1}, \quad (29)$$

$$\Delta(s, t) = \mathbf{X}' \text{diag}(F_0[\mathbf{x}_i' \beta(s \wedge t)] - F_0[\mathbf{x}_i' \beta(s)] F_0[\mathbf{x}_i' \beta(t)]) \mathbf{X}. \quad (30)$$

**Corollary 1** *Under the assumptions of the theorem,  $\hat{\mathbf{T}}(h, J)$  asymptotically has a multivariate normal distribution. Under  $H_0^{\text{NPar}}(F_0)$ , the covariance matrix of  $\mathbf{C}\hat{\mathbf{T}}(h, J)$  may be estimated consistently by*

$$\widehat{\text{Cov}}(\mathbf{C}\hat{\mathbf{T}}(h, J)) = \mathbf{C} \left[ \sum_{i,j} w_{ni} w_{nj} \hat{\mathbf{K}}(Y_{n:i}, Y_{n:j}) \right] \mathbf{C}',$$

where  $\hat{\mathbf{K}}(s, t)$  is obtained from (23), (29), (30), replacing  $\beta(t)$  with  $\hat{\beta}(t)$ .



## 5 Extension to nonlinear models and application to ROC analysis

In this section we briefly outline the extension of our method to nonlinear models and its application to the analysis of ROC (receiver operating characteristic) curves when covariables have to be taken into account.

Transformation models and ROC models are in close correspondence because of their invariance. For a given statistical model

$$Y_i \sim F_i,$$

the matrix  $\rho(\mathbf{F}) = \{\rho_{ij}\}$  of ROC functions

$$\rho_{ij} = F_i \circ F_j^{-1}$$

extracts the ordinal invariant part of information contained in that model. As at the level of *statistics* the vector  $\mathbf{R}(\mathbf{Y})$  of ranks is maximally invariant, at the level of *functionals*, the matrix of ROC functions has this property: Any ordinal invariant functional  $T^*(\mathbf{F})$  may be written as a functional of  $\rho(\mathbf{F})$ .

Parametric ROC models correspond to semiparametric transformation models. For example, the parametric ROC model

$$\rho_{ij}(u) = F_0(F_0^{-1}(u) + \mu_j - \mu_i)$$

corresponds to  $M^{\text{SPar}}(F_0)$  (3).

On the other hand, starting instead of (3) with the heteroscedastic transformation model

$$Y_i \sim F_0\left(\frac{a(t) - \mu_i}{\sigma_i}\right), \quad \mu_i = \boldsymbol{\beta}' \mathbf{x}_i, \quad \sigma_i = \boldsymbol{\gamma}' \mathbf{x}_i, \quad (31)$$

we arrive at the ROC model

$$\rho_{ij}(u) = F_0\left(\frac{\sigma_j}{\sigma_i} F_0^{-1}(u) + \frac{\mu_j - \mu_i}{\sigma_i}\right), \quad (32)$$

which e.g. was used as starting point in Hsieh (1996) (for  $F_0$  the extreme value c.d.f.).

Tosteson and Begg (1988) proposed to analyze the discrete-response version of (31) respective (32) with the PLUM-program of McCullagh (they assumed  $F_0$  to be the logistic c.d.f.). Considering the nonparametric version of (31) ( $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$  depending on  $t$ ) leads us to nonlinear nonparametric transformation model

$$\mathbb{P}(Y_i \leq t) = F_0 \circ g(\mathbf{x}_i, \boldsymbol{\vartheta}(t)), \quad \boldsymbol{\vartheta} : \mathbb{R} \rightarrow \Theta \quad \text{unknown, smooth,}$$

where

$$g : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$$

is a known, smooth second link function and  $\Theta$  a Euclidean parameter space. In model (31) we would use  $g(\mathbf{x}_i, \beta, \gamma) = \beta' \mathbf{x}_i / \gamma' \mathbf{x}_i$ . The corresponding family of estimating equations is

$$\sum_{j=1}^n F_0 \circ g(\mathbf{x}, \hat{\boldsymbol{\vartheta}}(t)) \mathbf{x}_j = \sum_{j=1}^n c(t - Y_j) \mathbf{x}_j,$$

which however in general is not the gradient condition of a convex function, hence some additional assumptions are necessary to guarantee consistent estimators.

In an analogous manner, we may define the nonlinear regression rank score process  $u \mapsto n^{1/2}(\hat{\boldsymbol{\vartheta}} \circ \hat{H}^{-1}(u) - \boldsymbol{\vartheta} \circ H^{-1}(u))$ , nonlinear regression rank statistics  $\hat{T}_x(h, J) = \int g(\mathbf{x}, \hat{\boldsymbol{\vartheta}}(t)) dJ \circ \hat{H}(t)$ , corresponding to functionals  $T_x(h, J) = \int g(\mathbf{x}, \boldsymbol{\vartheta}(t)) dJ \circ H(t)$ , hypotheses  $H_0 : g^*(\mathbf{x}, \boldsymbol{\vartheta}(t)) \equiv \mathbf{0}$  ( $g^*$  another known link function) and test statistics based on  $\hat{T}_x^*(h, J) = \int g^*(\mathbf{x}, \hat{\boldsymbol{\vartheta}}(t)) dJ \circ \hat{H}(t)$ .

## Acknowledgements

I would like to thank Roger Koenker for valuable hints to related literature.

## References

- [1] Adichie, J. N. (1984). Rank tests in linear models. In *Handbook of Statistics*, Vol.4, Eds. P.R. Krishnaiah and P. K. Sen, pp. 229-257.
- [2] Akritas M. G. and Arnold S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measurement designs. *J. Am. Statist. Assoc.* **89**, 336-343.
- [3] Anderson, J. A. and Philips, P. R. (1981). Regression discrimination and measurement models for ordered categorical variables. *Appl. Statist.* **30**, 22-31.
- [4] Bickel, P. J., Klaassen, C. A. J., Ritov Y. and Wellner J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. John Hopkins University Press.
- [5] Brunner E. and Puri M. L. (1996). Nonparametric methods in design and analysis of experiments. In *Handbook of Statistics*, Vol.13, Eds. C.R. Rao and S. Ghosh.
- [6] Chauduri P., Doksum K. and Samarov A. (1994). Regression estimators based on conditional quantiles. *J. Nonpar. Statist.* **3**, 317-321.

- [7] Cox D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc. B* **34**, 187–220.
- [8] Cuzick J. (1988). Rank regression. *Ann. Statist.* **16**, 1369–1389.
- [9] Doksum K. A. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *Ann. Statist.* **15**, 325–345.
- [10] Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. New York: Wiley.
- [11] Gutenbrunner, C. and Jurečková, J. (1992). Regression rank scores and regression quantiles. *Ann. Statist.* **20**, 305–330.
- [12] Hoeffding, W. (1951). “Optimum” nonparametric tests. *Second Berkeley symp. Math Statist. Probab.*, 83–92.
- [13] Hsieh, F. (1996). A transformation model for two survival curves: an empirical process approach. *Biometrika* **83**, 519–528.
- [14] Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.
- [15] Koenker, R. and Portnoy, S. (1987).  $L$ -estimation for linear models. *J. Amer. Statist. Assoc.* **82**, 851–857.
- [16] McCullagh P. (1980). Regression models for ordinal data. *J. R. Statist. Assoc. B* **42**, 109–127.
- [17] Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *J. R. Statist. Assoc. B* **44**, 234–243.
- [18] Pettitt, A. N. (1983). Approximate methods using ranks for regression with censored data. *Biometrika* **70**, 121–132.
- [19] Pettitt, A. N. (1987). Estimates for a regression parameter using ranks. *J. R. Statist. Assoc. B* **49**, 58–67.
- [20] Puri, M. L. and Sen, P. K. (1985). *Nonparametric Methods in General Linear Models*. New York: Wiley.
- [21] Pyke, R. and Shorak, G. (1968). Weak convergence of a two-sample empirical process and a new approach to Chernoff-Savage theorems. *Ann. Statist.* **39**, 755–771.
- [22] Serfling R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
- [23] Tosteson A. N. A. and Begg C. B. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making* **8**, 204–215.
- [24] Tsukahara, H. (1992). A rank estimator in the two-sample transformation model with randomly censored data. *Ann. Inst. Statist. Math.* **44**, 313–333.
- [25] Williams O. D. and Grizzle, J. E. (1972). Analysis of contingency tables having ordered categories. *J. Am. Statist. Assoc.* **67**, 55–63.

# Unimodality and the asymptotics of M-estimators

**Marc Hallin**

*Université Libre de Bruxelles, Belgium*

**Ivan Mizera**

*Comenius University, Bratislava, Slovakia*

*Abstract:* Unimodality, in its weaker and stronger forms, enters the robustness investigations somehow less often than symmetry. We point out how unimodality affects the asymptotics of M-estimators under heterogeneous (“non-i.i.d.”) errors. Sufficient conditions are given for consistency, with rates, of M-estimators in unimodal heterogeneous location models. For heteroscedastic models, a particular case of heterogeneous ones, a necessary and sufficient consistency condition, with rates, is provided for the  $L_1$  estimator - the sample median.

*Key words:* Sample heterogeneity, heterodasticity, unimodality, M-estimators, consistency.

AMS subject classification: Primary 62F35, 62F12; secondary 62F10, 60F05.

## 1 Introduction

In robustness theory, the assumption of symmetry is adopted quite regularly, although with a bit of strange taste: as pointed out by Huber (1981, page 95), “a restriction to exactly symmetric distributions . . . violates the very spirit of robustness”—since it is not stable under small perturbations of the underlying probabilities. On the other hand, the symmetry assumption resolves a dilemma of estimands—the problem of finding the target of location estimation. For symmetric population distributions, the center of symmetry is widely accepted as the “natural” location parameter—see, for instance, Hoaglin, Mosteller and Tukey (1983, chapter 9). And, needless to say, symmetry considerably simplifies a number technical considerations.

Unimodality, in its weaker and stronger forms, enters the robustness investigations somehow less often. It shares the instability of the symmetry—small perturbation (in weak topology sense) of a unimodal probability can result in a non-unimodal one. However, in terms of “realism”, unimodality performs better; recall only that almost all parent distributions involved in parametric models are unimodal (compared to a considerable number of asymmetric ones). And, even in the symmetric case, the center of symmetry frequently also is the mode. The impact of unimodality on the philosophy of estimation is perhaps not that unambiguous: it could be a matter of a discussion whether the mode, instead of the center of symmetry, can fulfill the need for the “natural” location in models with asymmetric (but unimodal) parent distribution. At least, the consequences of substituting unimodality assumptions for the symmetry ones deserve being closely investigated.

Skipping this problem, we shall concentrate on the technical virtues of unimodality. These are really rewarding. For instance, as shown in Mizera (1994), unimodality is most helpful in establishing the consistency of redescending M-estimators. The paper of Freedman and Diaconis (1982) pointed out that M-estimators can be inconsistent, due to non-identifiability—the lack of well-defined population value—unless the score function is monotone or the underlying distribution is symmetric and unimodal. Mizera (1994) showed that unimodality ensures the uniqueness of the population value also in asymmetric cases, for the majority of M-estimators with the non-monotone (“redescending”) score functions used in practice.

In this note, we point out how unimodality affects the asymptotic of M-estimators under heterogeneous errors. The violation of the i.i.d. assumption, the assumption which is central to most existing statistical models (recall i.i.d. error terms in regression or i.i.d. innovation process in time series) can arise from contamination, population heterogeneity, uncontrollable and hidden confounding factors, and variations in the measurement techniques or environmental conditions, the characteristics of which may vary through time and space. Different aspects of the asymptotics of M-estimators under heterogeneity were studied in Mizera and Wellner (1996) and Hallin and Mizera (1996). We concentrate here on the more specific consequences of unimodality. Sufficient conditions for consistency, with rates, are given for unimodal heterogeneous location models. For heteroscedastic models, a necessary and sufficient consistency condition, with rates, is established for the L1 estimator—the sample median. This condition is considerably more general than an earlier one by Sen (1968).

## 2 Consistency of M-estimators in heterogeneous location models

Associated with a nondecreasing score function  $\psi$  we define

$$\lambda_n(\psi, t) = \frac{1}{n} \sum_{i=1}^n \psi(X_{ni} - t).$$

An M-estimate is defined to be a “solution” of the equation  $\lambda_n(\psi, t) = 0$ ; since  $\lambda_n(\psi, t)$  is monotone, the values of  $t$  at which  $\lambda_n(\psi, t)$  crosses the zero level constitute an interval. To avoid ambiguity, we define the *M-estimate* to be the infimum of this interval:

$$T_{n,\psi} = \sup\{t : \lambda_n(\psi, t) > 0\}.$$

A *heterogeneous location model* consists of

- (H1) a set of data  $X_{n1}, X_{n2}, \dots, X_{nn}$ , which can be viewed as realizations of independent random variables
- (H2) with distribution functions  $F_{n1}, F_{n2}, \dots, F_{nn}$ , respectively,
- (H3) such that  $E[\psi(X_{ni} - \theta)] = 0$ , for all  $i = 1, 2, \dots, n$  and for all  $n = 1, 2, \dots$ .

The framework of (H1)–(H3) reduces to the standard i.i.d. one whenever  $F_{n1} = F_{n2} = \dots = F_{nn} = F_\theta$ , where  $F_\theta(x) = F(x - \theta)$ . In such a case, the Fisher consistency condition (H3) reduces to a much simpler and traditional one involving  $F_\theta$  only. However, in heterogeneous situation we need (H3) as the only thread connecting all  $X_{ni}$ ’s and  $F_{ni}$ ’s together, ensuring that our estimation of  $\theta$  makes any sense at all, that our data are not “a bizarre melange without much statistical relevance” (Le Cam 1986, page 529).

A heterogeneous location model is called *symmetric* if all  $F_{ni}$ ’s are symmetric about  $\theta$ , and *unimodal* if all  $F_{ni}$ ’s are unimodal with mode  $\theta$ ,  $i = 1, 2, \dots, n$ ,  $n = 1, 2, \dots$ . A distribution  $G$  is called unimodal (with mode  $\theta$ ) if it possesses a density  $g$  which is nondecreasing on  $(-\infty, \theta]$  and nonincreasing on  $[\theta, \infty)$ .

Various other unimodality concepts could have been considered. The weakest one, merely requiring the existence of a unique global maximum for the density, is too weak for most purposes. The slightly stronger definition formulated in terms of convexity and concavity of the distribution function  $G$  before and after the mode is closely related to ours—the only difference is that, unlike ours, it allows for an atom located at the mode. As for

the *strong unimodality* concepts (log-concavity, for instance), they are not needed here.

Note that, in symmetric models, (H3) automatically holds whenever  $\psi$  is odd (that is,  $\psi(-x) = -\psi(x)$ ). This, together with a natural interest in estimating the center of symmetry in the symmetric models explains the almost exclusive choice of odd score functions  $\psi$  in the practice of M-estimation. Thus, we shall assume about our score functions  $\psi$  that

(P1)  $\psi$  is a non-decreasing and odd function.

To ensure robustness, we adopt boundedness; since a multiple of  $\psi$  yields the same M-estimates, we set

(P2)  $\psi(-\infty) = -1$ ,  $\psi(\infty) = 1$ .

To avoid pathologies, we also suppose that

(P3) the set of discontinuity points of  $\psi$  is finite

and, finally, that

(P4)  $\psi$  is increasing at 0: for every  $\varepsilon > 0$ , there is a  $\delta(\varepsilon) > 0$  such that  $\psi(\varepsilon) - \psi(-\varepsilon) = 2\psi(\varepsilon) \geq 2\delta(\varepsilon)$ .

We remark that both (P3) and (P4) are satisfied by all score functions used in practice. For unimodal distributions, (P4) can ensure identifiability (uniqueness of the “population value”) of the M-estimator.

Consistency holds whenever the model is *conservative*: that is, whenever the sequence of average distribution functions

$$\bar{F}_n(x) = \frac{1}{n} \sum_{i=1}^n F_{ni}(x)$$

is tight (weakly sequentially compact; recall that a sequence  $G_n$  is tight if and only if for any  $\varepsilon > 0$  there is a  $K_\varepsilon > 0$  such that  $G_n(-K_\varepsilon) \leq \varepsilon/2$  and  $G_n(K_\varepsilon) \geq 1 - \varepsilon/2$  for all  $n$ ). An important special case of conservative model is the *mixture model*: the sequence  $\bar{F}_n$  converges weakly to a (proper) distribution function  $\bar{F}$ . The behavior of robust estimators in mixture models was studied by Stigler (1976).

**Theorem 1** *Suppose that  $X_{ni}$  satisfy the assumptions (H1)–(H3) of the heterogeneous location model; suppose further that this model is unimodal and conservative. If (P1)–(P4) hold, then  $T_{n,\psi} - \theta = o_P(1)$  as  $n \rightarrow \infty$ .*

**Proof:** See Section 3.

Under a slightly stronger assumption about  $\psi$

- (P4') there are non-negative integers  $q_1, q_2$  such that  $q_1 + q_2 = 1$ , and functions  $\psi_1, \psi_2$  satisfying (P1) and (P2) such that  $\psi = q_1\psi_1 + q_2\psi_2$ , where  $\psi_1(x) = \text{sign}(x)$  and  $\psi_2$  is absolutely continuous on some interval  $[-\Delta, \Delta]$ ,  $\Delta > 0$ , with a derivative  $\psi'$  satisfying  $\psi'(x) \geq K$  for all  $x \in [-\Delta, \Delta]$  — for some  $K > 0$

Theorem 1 can be strengthened to yield consistency rates.

**Theorem 2** *Suppose that  $X_{ni}$  satisfy the assumptions (H1)–(H3) of the heterogeneous location model; suppose further that this model is unimodal and conservative. If (P1)–(P3) and (P4') hold, then  $T_{n,\psi} - \theta = O_P(n^{-1/2})$  as  $n \rightarrow \infty$ .*

**Proof:** This theorem directly follows from Theorem 6 of Hallin and Mizera (1996).

Theorems 1 and 2 show that robust M-estimators behave in unimodal and conservative heterogeneous location models like in the i.i.d. case, where it can be said that they are always consistent—as soon as the corresponding population values are identifiable (see Huber 1981, page 54). The assumptions of Theorems 1 and 2 are easily checked in the particular case of *heteroscedastic location models*: heterogeneous location models with distribution functions satisfying

$$F_{ni}(x) = F\left(\frac{x - \theta}{c_{ni}}\right),$$

where  $c_{n1}, c_{n2}, \dots, c_{nn}$  are positive *scaling constants* and  $F$  is the distribution function of a fixed *parent distribution*. Note that every heteroscedastic model with symmetric and/or unimodal  $F$  is itself symmetric and/or unimodal.

For heteroscedastic models, we are able to state necessary and sufficient consistency conditions, with rates, for the special case of the L1 estimator—the sample median. Compared to the general conditions established in Mizera and Wellner (1996), our condition is specially tailored for heteroscedastic models, since, under a very mild regularity condition

- (S) the parent distribution admits a density  $f$  which is bounded, and there are  $\lambda > 0$ ,  $L > 0$  such that  $f(x) \geq L$  for  $x \in [-\lambda, \lambda]$ ,



which is clearly satisfied by any unimodal parent distribution with bounded density, it involves only the “empirical distribution” of the scaling constants. The conservativeness of the model is no longer required; our results hold, in particular, when some part of “probability mass” is allowed to “escape to infinity”. Let  $\Phi_c$  be the function from  $(0, \infty)$  to  $(0, \infty)$  defined by  $\Phi_c(x) = 1/c$  if  $x \leq c$  and  $1/x$  if  $x \geq c$ .

**Theorem 3** *Let  $T_{n,\psi}$  be the sample median (that is,  $\psi(x) = \text{sign}(x)$ ). If a heteroscedastic model satisfies (S), then  $T_{n,\psi} - \theta = o_P(r_n^{-1})$  if and only if*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \Phi_c(r_n c_{ni}) \rightarrow \infty \quad \text{as } n \rightarrow \infty, \quad (1)$$

for any fixed  $c > 0$ .

**Proof:** See Section 3.

Note that the choice of  $c$  for  $\Phi_c$  is inessential, due to the following elementary inequality, holding for any  $c \leq d$

$$\Phi_c(x) \geq \Phi_d(x) \geq \frac{c}{d} \Phi_c(x).$$

In other words, (1) holds for all  $c > 0$  as soon as it holds for one  $c > 0$ . Condition (1) implies that

$$\frac{1}{r_n \sqrt{n}} \sum_{i=1}^n \frac{1}{c_{ni}} \rightarrow \infty \quad \text{as } n \rightarrow \infty. \quad (2)$$

Under a non-degeneracy assumption, together with conservativeness and some additional regularity requirements, Sen (1968) proved an asymptotic normality result, from which (2) follows as a necessary and sufficient consistency condition. In fact, (1) and (2) are equivalent as soon as all  $c_{ni} \geq c$  for some  $c > 0$ . For the particular case of plain  $o_P(1)$  consistency ( $r_n = 1$ ), we obtain that (1) is equivalent to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \Phi_c(c_{ni}) \rightarrow \infty \quad \text{as } n \rightarrow \infty, \quad (3)$$

the condition already established in Hallin and Mizera (1996). Finally, Lemma 6 of Hallin and Mizera (1996) yields the following corollary.

**Theorem 4** *Under the assumptions of Theorem 3,  $T_{n,\psi} - \theta = O_P(s_n^{-1})$  if and only if (1) holds (for some  $c > 0$ ) for any sequence  $r_n$  such that  $r_n = o(s_n)$ .*

**Proof:** A direct consequence of Theorem 3 and Lemma 6 of Hallin and Mizera (1996).

### 3 Proofs

In the proofs, we write  $\theta = 0$ , without loss of generality.

**Lemma 1** *Suppose that the sequence  $\bar{F}_n$ ,  $n = 1, 2, \dots$  is tight and that the corresponding densities  $\bar{f}_n$  are unimodal with the common mode  $\theta$ . Then, for every  $\eta > 0$ , there is a  $K_\eta$  such that*

$$1 - \eta \leq \int_{-K_\eta}^{K_\eta} \bar{f}_n(x) dx \quad (4)$$

and

$$\max\{\bar{f}_n(-K_\eta), \bar{f}_n(K_\eta)\} \leq \eta \quad (5)$$

for all  $n = 1, 2, \dots$ .

**Proof:** First note that (4) is just the tightness condition rewritten in terms of densities. Turning to (5), assume that it does not hold: then, there is an  $\eta > 0$  such that for all  $K$

$$\text{either } \bar{f}_n(-K) > \eta \quad \text{or} \quad \bar{f}_n(K) > \eta. \quad (6)$$

Set  $K = 2/\eta$  and suppose that  $\bar{f}_n(-K) > \eta$ , say. By unimodality,

$$\int_{-K}^0 \bar{f}_n(x) dx \leq \int_{-K}^0 \eta dx = 2$$

— a contradiction; the other case in (6) is treated analogously. At this point, we could possibly have one value of  $K$  for (4) and another one for (5) — but the maximum of them two works at both.

**Proof of Theorem 1:** For any  $\varepsilon > 0$ , let

$$a_n(\psi, \varepsilon) = E\lambda_n(\psi, \theta - \varepsilon) = \frac{1}{n} \sum_{i=1}^n \int \psi(x - \theta + \varepsilon) dF_{ni}(x)$$

and

$$b_n(\psi, \varepsilon) = E\lambda_n(\psi, \theta + \varepsilon) = \frac{1}{n} \sum_{i=1}^n \int \psi(x - \theta - \varepsilon) dF_{ni}(x).$$

In view of Theorem 1 of Hallin and Mizera (1996), it is sufficient to show that for any  $\varepsilon > 0$ ,  $a_n(\psi, \varepsilon)$  and  $b_n(\psi, \varepsilon)$  are bounded away from zero for sufficiently large  $n$ . We give a proof for  $a_n(\psi, \varepsilon)$ ; the proof for  $b_n(\psi, \varepsilon)$  is entirely similar. In view of (H3), it is sufficient to show that, for any  $\varepsilon > 0$  (setting again  $\theta = 0$ ),

$$\int [\psi(x + \varepsilon) - \psi(x)] \bar{f}_n(x) dx > 0 \quad (7)$$

for sufficiently large  $n$  (note that, due to monotonicity, the integrand in (7) is non-negative).

Fix  $\varepsilon > 0$ . By (P4), there is a  $\delta(\varepsilon/2)$  such that  $\psi(\varepsilon/2) \geq \delta(\varepsilon/2)$ ; hence we have, for all  $x \in [-\varepsilon/2, 0]$ ,

$$\psi(x + \varepsilon) - \psi(x) \geq \psi(x + \varepsilon) \geq \psi(\tfrac{1}{2}\varepsilon) \geq \delta(\tfrac{1}{2}\varepsilon). \quad (8)$$

Now, choose  $\eta$  and  $C > 0$  such that for  $K_\eta$  given by Lemma 1 we have

$$\eta + CK_\eta \leq \min\{\tfrac{1}{2}\delta(\tfrac{1}{2}\varepsilon), \tfrac{1}{4}\}.$$

If  $\bar{f}_n(\varepsilon/2) \geq C$ , we have by (8)

$$\int [\psi(x + \varepsilon) - \psi(x)] \bar{f}_n(x) dx \geq \int_{-\varepsilon/2}^0 \delta(\tfrac{1}{2}\varepsilon) \bar{f}_n(x) dx \geq \tfrac{1}{2}\varepsilon C \delta(\tfrac{1}{2}\varepsilon),$$

due to unimodality. If  $\bar{f}_n(\varepsilon/2) \leq C$ , Lemma 1 gives

$$\begin{aligned} \int_{-\infty}^{-\varepsilon/2} \bar{f}_n(x) dx &= \int_{-\infty}^{-K_\eta} \bar{f}_n(x) dx + \int_{-K_\eta}^{-\varepsilon/2} \bar{f}_n(x) dx \\ &\leq \eta + (K_\eta - \tfrac{1}{2}\varepsilon)C \leq \eta + CK_\eta \end{aligned}$$

due to unimodality again; hence,

$$\int_{-\varepsilon/2}^{\infty} \bar{f}_n(x) dx \geq 1 - \eta - CK_\eta$$

and, consequently,

$$\begin{aligned} \int [\psi(x + \varepsilon) - \psi(x)] \bar{f}_n(x) dx &= \int \psi(x + \varepsilon) \bar{f}_n(x) dx \\ &= \int_{-\infty}^{-\varepsilon/2} \psi(x + \varepsilon) \bar{f}_n(x) dx + \int_{-\varepsilon/2}^{\infty} \psi(x + \varepsilon) \bar{f}_n(x) dx \\ &\geq \int_{-\infty}^{-\varepsilon/2} -1 \bar{f}_n(x) dx + \int_{-\varepsilon/2}^{\infty} \delta(\tfrac{1}{2}\varepsilon) \bar{f}_n(x) dx \\ &\geq -\eta - CK_\eta + \delta(\tfrac{1}{2}\varepsilon)(1 - \eta - CK_\eta) \\ &\geq -\tfrac{1}{2}\delta(\tfrac{1}{2}\varepsilon) + \tfrac{3}{4}\delta(\tfrac{1}{2}\varepsilon) = \tfrac{1}{4}\delta(\tfrac{1}{2}\varepsilon). \end{aligned}$$

In both cases, we have that (7) is bounded from below by

$$\min\{\tfrac{1}{2}\varepsilon C \delta(\tfrac{1}{2}\varepsilon), \tfrac{1}{4}\delta(\tfrac{1}{2}\varepsilon)\} > 0,$$

which proves the statement.

**Proof of Theorem 3:** Let  $\psi(x) = \text{sign}(x)$ , let  $f$  be a density of the parent distribution of a heteroscedastic model satisfying (S). Then,  $f$  is bounded by  $K$ ; without loss of generality we may suppose that  $K \geq 1$ .

*Necessity.* By Theorem 3 of Hallin and Mizera (1996),  $o_P(r_n^{-1})$  consistency implies that  $\sqrt{n}a_n(\psi, r_n^{-1}) \rightarrow \infty$  as  $n \rightarrow \infty$ . Proceeding similarly as in the proof of Theorem 8 of Hallin and Mizera (1996), we obtain

$$\begin{aligned}
 \sqrt{n}a_n(\psi, r_n^{-1}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_{-(r_n c_{ni})^{-1}}^0 f(y) dy \\
 &= \frac{1}{\sqrt{n}} \left[ \sum_{r_n c_{ni} \leq 1} \int_{-(r_n c_{ni})^{-1}}^0 f(y) dy + \sum_{r_n c_{ni} > 1} \int_{-(r_n c_{ni})^{-1}}^0 f(y) dy \right] \\
 &\leq K \frac{1}{\sqrt{n}} \left[ \sum_{r_n c_{ni} \leq 1} 1 + \sum_{r_n c_{ni} > 1} \frac{1}{r_n c_{ni}} \right] \\
 &= K \frac{1}{\sqrt{n}} \sum_{i=1}^n \Phi_1(r_n c_{ni}),
 \end{aligned}$$

and (1) follows.

*Sufficiency.* Let (S) hold with  $\lambda$  and  $L$ ; choose  $\Gamma$  and  $c$  such that  $\Gamma = \lambda c$ . If  $0 \leq \varepsilon \leq \Gamma$ , then, as in the proof of Theorem 7 of Hallin and Mizera (1996),

$$\begin{aligned}
 a_n(\psi, \varepsilon r_n^{-1}) &= \frac{1}{n} \sum_{i=1}^n \int_{-\varepsilon(r_n c_{ni})^{-1}}^0 f(y) dy \\
 &= \frac{1}{n} \left[ \sum_{r_n c_{ni} \leq c} \int_{-\varepsilon(r_n c_{ni})^{-1}}^0 f(y) dy + \sum_{r_n c_{ni} > c} \int_{-\varepsilon(r_n c_{ni})^{-1}}^0 f(y) dy \right] \\
 &\geq \frac{1}{n} \left[ \sum_{r_n c_{ni} \leq c} \int_{-\varepsilon c^{-1}}^0 f(y) dy + \sum_{r_n c_{ni} > c} \int_{-\varepsilon(r_n c_{ni})^{-1}}^0 f(y) dy \right] \\
 &\geq L\varepsilon \frac{1}{n} \left[ \sum_{r_n c_{ni} \leq c} \frac{1}{c} + \sum_{r_n c_{ni} > c} \frac{1}{r_n c_{ni}} \right] \\
 &= L\varepsilon \frac{1}{n} \sum_{i=1}^n \Phi_c(r_n c_{ni}).
 \end{aligned}$$

An entirely similar argument for  $b_n(\psi, \varepsilon r_n^{-1})$  and the subsequent application—assuming (1)—of Theorems 2 and 3 of Hallin and Mizera (1996) conclude the proof.

## Acknowledgements

For the first author the research was supported by the Fonds d'Encouragement à la Recherche de l'Université Libre de Bruxelles and the European Human Capital contract ERB CT CHRX 940963 and for the second author, the research was supported by the Fonds National de la Recherche Scientifique, the Banque Nationale de Belgique, and Slovak GAS grant 1/1489/94.

## References

- [1] Freedman, D. A. and Diaconis, P. (1982). On inconsistent M-estimators. *Ann. Statist.* **10**, 454–461.
- [2] Hallin, M. and Mizera, I. (1995). Sample heterogeneity and the asymptotics of M-estimators. Preprint IS-P 1996-15 (No. 49), Institut de Statistique de l'Université Libre de Bruxelles, Brussels.
- [3] Hoaglin D. C., Mosteller F. M., and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.
- [4] Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.
- [5] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Wiley.
- [6] Mizera, I. (1994) On consistent M-estimators: tuning constants, unimodality and breakdown. *Kybernetika* **30**, 289–300.
- [7] Mizera, I. and Wellner, J. A. (1996). Necessary and sufficient conditions for the consistency of the sample median of independent but not identically distributed random variables. Preprint IS-P 1996-6 (No. 40), Institut de Statistique de l'Université Libre de Bruxelles, Brussels.
- [8] Sen, P. K. (1968). Asymptotic normality of sample quantiles for  $m$ -dependent processes. *Ann. Math. Statist.* **39**, 1724–1730.
- [9] Stigler, S. M. (1976). The effect of sample heterogeneity on linear functions of order statistics, with applications to robust estimation. *J. Amer. Statist. Assoc.* **71**, 956–960.

# *L<sub>1</sub>-test procedures for detection of change*

M. Hušková

*Charles University, Prague, Czech Republic*

*Abstract:* *L<sub>1</sub>-type test procedures for detection of a change in linear models are proposed, their properties are studied under the null hypothesis (no change).*

*Key words:* *L<sub>1</sub>-test procedures, linear models, change point.*

AMS subject classification: 62G20, 62E20.

## 1 Introduction

The problem to detect and to identify changes in statistical models has attracted a number of researchers in the last two decades. Using various principles they have proposed a number of statistical procedures that are sensitive w.r.t. detection of changes, have studied their (mostly) limit properties and, also, have applied to real data sets.

The problem of detection and identification of changes in statistical models is known as the *change point problem* (mostly for case of changes in location models), *disorder problem* or *testing the constancy of regression relationship over time*. These problems arise in a number of applications (economic modelling, quality control, biology, medicine, meteorology and ecology among others).

We shall consider here the following regression model with possible change after an unknown time point  $m$ :

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{x}_i \delta_n I\{i > m\} + E_i, \quad i = 1, \dots, n, \quad (1)$$

where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ ,  $x_{i1} = 1$ ,  $i = 1, \dots, n$ , are known regression vectors,  $m(< n)$ ,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ ,  $\delta_n = (\delta_{n1}, \dots, \delta_{np})^T$  are unknown parameters,  $E_1, \dots, E_n$  are i.i.d. random variables with common distribution function  $F$ .  $I\{A\}$  denotes the indicator of the set  $A$ .

The model corresponds to the situation when up to an unknown  $m$  the observations follow the regression model with the regression parameter  $\beta$  and then the model changes to the regression model with the regression parameter  $\beta + \delta_n$ . The parameter  $m$  is called the *change point*.

The problem of our interest is to test

$$H_0 : m = n \text{ against } H_1 : m < n.$$

The authors usually apply either the likelihood ratio principle or the Bayesian approach. The first principle leads to max-type procedures the other gives sum-type procedures.

First, we shall describe likelihood ratio and related procedures when the distribution of the error terms  $F$  is  $N(0, \sigma^2)$ ,  $\sigma^2 > 0$  known. It will give motivation how to develop  $L_1$ -procedures.

Assume that  $e'_i$ s are i.i.d. with distribution  $N(0, \sigma^2)$ ,  $\sigma^2 > 0$  known, the likelihood ratio principle leads to the test statistics

$$T_{n,LSE} = \max_{p < k < n-p} \left\{ - \sum_{i=1}^k \rho_{LSE}(Y_i - x_i^T \beta_{k,LSE}) - \sum_{i=k+1}^n \rho_{LSE}(Y_i - x_i^T \beta_{k,LSE}^*) + \sum_{i=1}^n \rho_{LSE}(Y_i - x_i^T \beta_{n,LSE}) \right\} / \sigma^2, \quad (2)$$

where  $\rho_{LSE}(x) = x^2$ ,  $x \in R^1$ ,  $\beta_{k,LSE}$  and  $\beta_{k,LSE}^*$  are the least squares estimators of the regression parameters based on  $X_1, \dots, X_k$  and  $X_{k+1}, \dots, X_n$ , respectively, i.e.,

$$\beta_{k,LSE} = C_k^{-1} \sum_{i=1}^k x_i Y_i, \quad \beta_{k,LSE}^* = C_k^{*-1} \sum_{i=k+1}^n x_i Y_i$$

with

$$C_k = \sum_{i=1}^k x_i x_i^T, \quad C_k^* = \sum_{i=k+1}^n x_i x_i^T. \quad (3)$$

The test statistics  $T_{n,LSE}$  can be expressed equivalently as:

$$T_{n,LSE} = \max_{p < k < n-p} \left\{ \left( \beta_{k,LSE} - \beta_{k,LSE}^* \right)^T (C_k^{-1} + C_k^{*-1})^{-1} \left( \beta_{k,LSE} - \beta_{k,LSE}^* \right) / \sigma^2 \right\} \quad (4)$$

and

$$T_{n,LSE} = \max_{p < k < n-p} \left\{ S_{k,LSE}^T (C_k^{-1} C_n C_k^{*-1}) S_{k,LSE} / \sigma^2 \right\}, \quad (5)$$

where

$$\mathbf{S}_{k,LSE} = \sum_{i=1}^k \mathbf{x}_i (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{n,LSE}). \quad (6)$$

Horváth (1995) among others derived the limit distribution of  $T_{n,LSE}$  under  $H_0$  and showed that if mild assumptions are satisfied then under  $H_0$ , as  $n \rightarrow \infty$ ,

$$\max \left\{ \mathbf{S}_{k,LSE}^T (\mathbf{C}_k^{-1} \mathbf{C}_n \mathbf{C}_k^{*-1}) \mathbf{S}_{k,LSE} / \sigma^2, \right. \\ \left. k = [n/\log n], \dots, n - [n/\log n] \right\} / T_{n,LSE} \rightarrow 0.$$

which means that asymptotically even under  $H_0$  the terms with  $k$  "small" or close to  $n$  dominate the others. To avoid to this unpleasant property some modifications were proposed. Namely, the class of test statistics depending on a suitable weight function  $q$  was introduced:

$$T_{n,LSE}(q) = \max_{1 \leq k < n} \left\{ \frac{\mathbf{S}_{k,LSE}^T \mathbf{C}_n^{-1} \mathbf{S}_{k,LSE}}{q^2(k/n) \sigma^2} \right\}. \quad (7)$$

Typical choices of the weight function  $q$  are the following

$$q(t) = (t(1-t))^{-1/2} \quad t \in (a_1, a_2) \quad (8)$$

$$q(t) = 0 \quad \text{otherwise,}$$

where  $0 < a_1 < a_2 < 1$ , or

$$q(t) = (t(1-t))^{-\gamma} \quad t \in (0, 1/2), \quad (9)$$

with  $\gamma \in [0, 1/2)$ .

Some authors (e.g. Jandhyala and MacNeill, 1989; Ploberger and Krämer, 1992) suggested to apply procedures based on properly standardized partial sums of the  $LSE$ -residuals:

$$S_{k,LSE}^o = \sum_{i=1}^k (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{n,LSE}), \quad k = 1, 2, \dots, n \quad (10)$$

They proposed a computationally feasible procedure:

$$T_{n,LSE}^o(q) = \max_{1 \leq k < n} \left\{ \frac{|S_{k,LSE}^o|}{\sqrt{n} \sigma q(k/n)} \right\}, \quad (11)$$

where  $q$  is a weight function. Another type of procedures is based on moving sums (MOSUM) of the LSE-residuals. They are defined by:

$$T_{n,LSE}^*(G) = \max_{G < k < n} \left\{ \frac{1}{\sqrt{G}} |S_{k,LSE}^o - S_{k-G,LSE}^o| / \sigma \right\} \quad (12)$$



and

$$T_{n,LSE}^{**} = \max_{G < k < n-G} \left\{ \frac{1}{\sqrt{G}} |S_{k+G,LSE}^o - 2S_{k,LSE}^o + S_{k-G,LSE}^o| / \sigma \right\} \quad (13)$$

Bayesian type of test statistics have the form

$$T_{n,LSE}^B(v) = \sum_{k=1}^{n-1} v(k/n) \left\{ \mathbf{S}_{k,LSE}^T \mathbf{C}_n^{-1} \mathbf{S}_{k,LSE} / \sigma^2 \right\}^{1/2}. \quad (14)$$

where  $v(1/n), \dots, v((n-1)/n)$  represent priors.

In spite that the procedures were developed for normally distributed random errors they can be applied also for nonnormally distributed random errors with zero mean and finite absolute moment of the order  $2+\Delta$  ( $\Delta > 0$ ) (in some cases a finite second moment suffices).

If  $\sigma^2$  is unknown it is recommended to estimate it by

$$\hat{\sigma}_{n,LSE}^2 = \frac{1}{n-p} \min_{1 \leq k \leq m} \left\{ \sum_{i=1}^k (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{k,LSE})^2 + \sum_{i=k+1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{k,LSE}^*)^2 \right\}.$$

and plug into the above statistics.

Typically large values of the introduced test statistics indicate that the null hypothesis  $H_o$  fails. The exact distributions even under  $H_o$  of the above introduced test statistics are unknown. The limit distributions were derived under mild assumptions on the distribution of the error terms  $E_i'$ s and on the design points  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , which enable to get the approximations for the critical values.

The test procedures corresponding to  $T_{n,LSE}$  and  $T_{n,LSE}(q)$  were studied by a number of authors, e.g. Quandt (1958, 1960), Worsley (1983), Kim and Siegmund (1989), Gombay and Horváth (1994), Horváth (1995), Antoch and Hušková (1992). Well known is the paper by Brown, Durbin and Evans (1975) devoted to the procedures based on recursive residuals. Bayesian type procedures were proposed and studied by Broemling and Tsurumi (1987) and Jandhyala and MacNeill (1989, 1991, 1992). Procedures based on partial sums of *LSE*-residuals were investigated, e.g., by Jandhyala and MacNeill (1989) and Ploberger and Krämer (1992). Hackl (1980) deeply studied procedures based on moving sums (MOSUM). A number of applications in econometrics is contained in Hackl (1989) and Hackl and Westlund (1991). Horváth, Hušková and Serbinowska (1995) considered the case when the change can occur in the regression parameters and/or in the scale  $\sigma$ .

Along the same line  $M$ -type and  $R$ - (rank based) tests were developed and studied. Some results on the  $M$ -procedures for changes in regression

models can be found, e.g., in Sen (1984) and Hušková (1990a,b, 1994 a,b).  $R$ -type test procedures were studied by Sen (1980, 1982) and Hušková (1994b).

## 2 $L_1$ procedures

It is easily seen that the test statistics  $T_{n,LSE}$ 's are functions of least squares estimators and of the  $LSE$ -residuals  $Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{LSE}$ , where  $\boldsymbol{\beta}_{LSE}$  is a least squares estimator, therefore  $T_{n,LSE}$ 's can be viewed as the  $L_2$ -type test statistic.

Now, along this line the  $L_1$ -procedures will be developed. Namely, we replace the  $LSE$  estimators by  $L_1$  estimators,  $\rho_{LSE}$  by  $\rho_{L_1}(x) = |x|$ ,  $x \in R^1$ , and the  $LSE$ -residuals by  $L_1$ -residuals  $\psi_{L_1}(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{L_1})$ , where  $\boldsymbol{\beta}_{L_1}$  is an  $L_1$  estimator  $\boldsymbol{\beta}$  and  $\psi_{L_1}(x) = -1$ ,  $x < 0$ ,  $\psi_{L_1}(x) = 0$ ,  $x = 0$ ,  $\psi_{L_1}(x) = 1$ ,  $x > 0$ .

From three equivalent expressions for  $T_{n,LSE}$  ((2), (4), (5)) we get three different test statistics. Namely,

$$T_{n,L_1}^{(1)} = \max_{p < k < n-p} \left\{ 2\hat{f}(F^{-1}(1/2)) \left( - \sum_{i=1}^k |Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{k,L_1}| \right. \right. \quad (15)$$

$$\left. \left. - \sum_{i=k+1}^n |Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{k,L_1}^*| + \sum_{i=1}^n |Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{n,L_1}| \right) \right\},$$

$$T_{n,L_1}^{(2)} = \max_{p < k < n-p} \left\{ 4\hat{f}^2(F^{-1}(1/2)) \left( \boldsymbol{\beta}_{k,L_1} - \boldsymbol{\beta}_{k,L_1}^* \right)^T \right. \quad (16)$$

$$\left. \left( \mathbf{C}_k^{-1} + \mathbf{C}_k^{*-1} \right)^{-1} \left( \boldsymbol{\beta}_{k,L_2} - \boldsymbol{\beta}_{k,L_1}^* \right) \right\}$$

and

$$T_{n,L_1}^{(3)} = \max_{p < k < n-p} \left\{ \mathbf{S}_{k,L_1}^T (\mathbf{C}_k^{-1} \mathbf{C}_n \mathbf{C}_k^{*-1}) \mathbf{S}_{k,L_1} \right\}, \quad (17)$$

where  $\hat{f}(F^{-1}(1/2))$  is an estimator of  $f(F^{-1}(1/2))$ ,  $F^{-1}$  and  $f$  denote the quantile function and the density, respectively,

$$\mathbf{S}_{k,L_1} = \sum_{i=1}^k \mathbf{x}_i \psi_{L_1}(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{n,L_1}), \quad (18)$$

and  $\boldsymbol{\beta}_{k,L_1}$  and  $\boldsymbol{\beta}_{k,L_1}^*$  are the  $L_1$ -estimators of the regression parameters based on  $X_1, \dots, X_k$  and  $X_{k+1}, \dots, X_n$ , respectively, i.e., they are defined as solutions of the minimization problems

$$\min \sum_{i=1}^k |Y_i - \mathbf{x}_i^T \mathbf{t}|, \mathbf{t} \in R^p \quad (19)$$

and

$$\min \sum_{i=k+1}^n |Y_i - x_i^T \mathbf{v}|, \mathbf{v} \in R^p, \quad (20)$$

respectively.

The statistic  $T_{n,L_1}^{(1)}$  is a likelihood ratio type statistic,  $T_{n,L_1}^{(2)}$  is a Wald type test statistic and  $T_{n,L_1}^{(3)}$  is a score type test statistic.

Computational feasibility of the statistic  $T_{n,L_1}^{(3)}$  is evident. The statistics  $T_{n,L_1}^{(1)}, T_{n,L_1}^{(2)}$  depend on the estimator of  $f(F^{-1}(1/2))$  and also on the estimators  $\beta_{k,L_1}$  and  $\beta_{k,L_1}^*$ ,  $k = 1, \dots, n$ . Quality of the estimator of  $f(F^{-1}(1/2))$  strongly influence the quality of the test itself.

The weighted type test statistics are defined by

$$T_{n,L_1}(q) = \max_{1 \leq k < n} \left\{ \frac{\mathbf{s}_{k,L_1}^T \mathbf{C}_n^{-1} \mathbf{s}_{k,L_1}}{q(k/n)} \right\}, \quad (21)$$

where the weight function  $q$  is the same as in  $LSE$ -case.

Next, we introduce the test statistics based on partial sums of  $L_1$ -residuals

$$S_{k,L_1}^o = \sum_{i=1}^k \psi_{L_1}(Y_i - \mathbf{x}_i^T \beta_{n,L_1}), \quad k = 1, 2, \dots, n \quad (22)$$

We get the weighted sum type and MOSUM type test statistics

$$T_{n,L_1}^o(q) = \max_{1 \leq k < n} \left\{ \frac{|S_{k,L_1}^o|}{\sqrt{nq(k/n)}} \right\}, \quad (23)$$

$$T_{n,L_1}^*(G) = \max_{G < k < n} \left\{ \frac{1}{\sqrt{G}} |S_{k,L_1}^o - S_{k-G,L_1}^o| \right\}, \quad (24)$$

$$T_{n,L_1}^{**} = \max_{G < k < n-G} \left\{ \frac{1}{\sqrt{2G}} |S_{k+G,L_1}^o - 2S_{k,L_1}^o + S_{k-G,L_1}^o| \right\}, \quad (25)$$

where  $q$  is a weight function.

Finally, Bayesian type of test statistics have the form

$$T_{n,L_1}^B(v) = \sum_{k=1}^{n-1} v(k/n) \left\{ 2\mathbf{s}_{k,L_1}^T \mathbf{C}_n^{-1} \mathbf{s}_{k,L_1} \right\}^{1/2}.$$

where  $v(1/n), \dots, v((n-1)/n)$  represent priors.

Analogously as in the  $L_2$  situation large values indicate that  $H_o$  does not hold. The exact distribution of the test statistics even under the null hypothesis can be hardly obtained. The limit distributions under  $H_o$  can be derived (see Theorem 1 - Theorem 3 below) which are then useful in getting approximations to the critical values.

Now, we pay attention to the limit behavior of the introduced tests statistics under the null hypothesis.

We consider the assumptions:

(i) Random variables  $Y_1, \dots, Y_n$  follow the model (1) with  $n = m$  and the distribution  $F$  has median 0 and Lipschitz of order  $\gamma_1 \in (0, 1]$  and strictly positive density at the median, i.e.,  $F^{-1}(1/2) = 0$ ,  $|f(0) - f(x)| \leq D|x|^{\gamma_1}$  for some  $D > 0$  and all  $x$  in a neighbourhood of 0 and  $f(0) > 0$ .

(ii)  $\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{C}_{[nt]} = t\mathbf{C}$ ,  $t \in [0, 1]$ , for some  $\mathbf{C} > 0$ .

(iii) There exist  $\epsilon \in (0, 1)$  and  $\gamma_2 > 0$  such that, as  $n \rightarrow \infty$ ,

$$\left\| \frac{1}{k} \mathbf{C}_k - \mathbf{C} \right\| = O(k^{-\gamma_2})$$

$$\left\| \frac{1}{n-k} \mathbf{C}_k^* - \mathbf{C} \right\| = O((n-k)^{-\gamma_2})$$

uniformly for  $1 \leq k \leq n\epsilon$ , where  $\mathbf{C}$  is the same as in (ii) and  $\|\cdot\|$  denotes the Euclidian norm.

(iv) As,  $n \rightarrow \infty$ ,

$$\max_{1 \leq k < n} \left\{ \frac{1}{k} \sum_{i=1}^k \|\mathbf{x}_i\|^3 + \frac{1}{n-k} \sum_{i=k+1}^n \|\mathbf{x}_i\|^3 \right\} = O(1).$$

(v)  $\hat{f}(0)$  be an estimator of  $f(0)$  such that, as  $n \rightarrow \infty$ ,

$$\hat{f}(0) - f(0) = o_p((\log \log n)^{-1/2}).$$

**Theorem 1** *Let assumptions (i), (ii), (iii) and (iv) be satisfied then*

$$\lim_{n \rightarrow \infty} P(a(\log n)(T_{n,L_1}^{(3)})^{1/2} \leq t + b_p(\log n)) = \exp\{-2 \exp\{-t\}\}, t \in R^1, \quad (26)$$

where

$$a(y) = (2 \log y)^{1/2}, \quad b_p(y) = 2 \log y + \frac{p}{2} \log \log y - \log(2\Gamma(p/2)), y > 1, \quad (27)$$

and

$$\Gamma(z) = \int_0^\infty t^{z-1} \exp\{-t\} dt.$$

If, moreover, (v) is satisfied then the assertion (26) remains true if  $(T_{n,L_1}^{(3)})^{1/2}$  is replaced by  $(T_{n,L_1}^{(1)})^{1/2}$  or  $(T_{n,L_1}^{(2)})^{1/2}$ .

**Theorem 2** *Let assumptions (i), (ii) and (iv) be satisfied then, as  $n \rightarrow \infty$ ,*

$$(T_{n,L_1}(q))^{1/2} \rightarrow^D \sup_{0 < t < 1} \left\{ \frac{(\sum_{i=1}^p B_i^2(t))^{1/2}}{q(k/n)} \right\}, \quad (28)$$

and

$$T_{n,L_1}^o(q) \rightarrow^D \sup_{0 < t < 1} \left\{ \frac{|B_1(t)|}{q(k/n)} \right\}, \quad (29)$$

$$T_{n,L_1}^B(v) \rightarrow^D \int_0^1 v(t) \left( \sum_{i=1}^p B_i^2(t) \right)^{1/2} dt, \quad (30)$$

where  $\{B_j(t); t \in (0, 1)\}$  are independent Brownian bridges and  $q$  is the weight function defined by either (8) or (9) and  $v(t) = (q(t)\sqrt{t(1-t)})^{-1}$ ,  $t \in (0, 1)$ .

**Theorem 3** Let assumptions (i), (ii) and (iv) be satisfied and let, as  $n \rightarrow \infty$ ,

$$G/n \rightarrow 0, \quad G^{-1}n^{2/3} \log n \rightarrow 0, \quad (31)$$

then

$$\lim_{n \rightarrow \infty} P\left(a(\log(n/G))T_{n,L_1}^*(G) \leq t + b_1(\log(n/G) + \log 2)\right) \quad (32)$$

$$= \exp\{-2 \exp\{-t\}\}, \quad t \in R^1,$$

$$\lim_{n \rightarrow \infty} P\left(a(\log(n/G))T_{n,L_1}^{**}(G) \leq t + b_1(\log(n/G) + \log 3)\right) \quad (33)$$

$$= \exp\{-2 \exp\{-t\}\}, \quad t \in R^1.$$

The assertions of Theorem 1-3 remain true if the  $L_1$ -test statistics are replaced by *LSE*-test statistics and if in assumption (i) the request of zero median is replaced by the request of zero mean and finite absolute moment of order  $2 + \Delta$ ,  $\Delta > 0$ .

Assertions (26), (32) and (33) are extreme value type theorems. It is known that the convergence in (26), (32) and (33) is rather slow.

The explicit form of the limit distribution in (28), (29) and (30) is known only for some weight function  $q$ , e.g., for  $q$  in (9) with  $\gamma = 0$  in Sen (1981) and for  $q$  in (8) Siegmund (1987) derived a proper approximation.

Approximation to the critical values corresponding to  $T_{n,L_1}^{(j)}$ ,  $j = 1, 2, 3$ ,  $T_{n,L_1}^*$  and  $T_{n,L_1}^{**}$  can be easily calculated using a pocket calculator.

The tests based on either of  $T_{n,L_1}^{(j)}$ ,  $j = 1, 2, 3$ ,  $T_{n,L_1}^B$  are consistent for fixed and as well as some local alternatives. Concerning  $T_{n,L_1}(q)$ ,  $T_{n,L_1}^o(q)$ ,  $T_{n,L_1}^*$  and  $T_{n,L_1}^{**}$  their limit distribution depend on  $\delta_n$  and the design matrix. This will be studied in a different paper.

The assumptions (ii) - (iv) imposed on the design matrix are slightly stronger than one usually assumes when studying for example  $L_1$  estimators in the model (1) with  $\delta_n = 0$ .

Concerning the estimator of  $f(0)$  we need an estimator that behaves reasonably well not only under the null hypothesis but also under alternatives. Such estimators can be described as follows:

$$\hat{f}(0) = \frac{\hat{m}}{n} \hat{f}_{\hat{m}}(0) + \frac{n - \hat{m}}{n} \hat{f}_{\hat{m}}^*(0),$$

where  $\hat{f}_{\hat{m}}(0)$  and  $\hat{f}_{\hat{m}}^*(0)$  are estimators of  $f(0)$  based on  $Y_1, \dots, Y_{\hat{m}}$  and  $Y_{\hat{m}+1}, \dots, Y_n$  and  $\hat{m}$  is an estimator of possible change point  $m$ . There is a number of possibilities how to estimate  $\hat{f}_{\hat{m}}(0)$ ,  $\hat{f}_{\hat{m}}^*(0)$  and  $\hat{m}$ . Here is one suggestion

$$\hat{m} = \operatorname{argmax}\{||\beta_{k,L_1} - \beta_{k,L_1}^*||; k = 1, \dots, n\},$$

$$\hat{f}_{\hat{m}}(0) = \frac{1}{2\eta\hat{m}^{1/2}} \sum_{i=1}^{\hat{m}} I\{\mathbf{x}_i^T \beta_{\hat{m},L_1} - \hat{m}^{-1/2}\eta \leq Y_i \leq \mathbf{x}_i^T \beta_{\hat{m},L_1} + \hat{m}^{-1/2}\eta\},$$

where  $\eta > 0$  fixed, and  $\hat{f}_{\hat{m}}^*(0)$  is defined accordingly. Under the assumptions (i),(ii) and (iv) the resulting estimator  $\hat{f}(0)$  has the property requested in the assumption (v).

### 3 Proofs

Since the proofs are quite technical we give only a sketch of them. First, we formulate several technical lemmas that are modifications of results proved elsewhere.

**Lemma 1** *Let assumptions (i) - (iv) be satisfied then for any  $\eta > 0$  there exist  $A_\eta > 0$  and  $n_\eta$  such that for all  $n \geq n_\eta$*

$$\begin{aligned} P\left(\sup\left\{\left|\sum_{i=1}^n (\rho_{L_1}(E_i - n^{-1/2}\mathbf{x}_i^T \mathbf{t}) - \rho_{L_1}(E_i) + n^{-1/2}\mathbf{x}_i^T \mathbf{t} \psi_{L_1}(E_i - n^{-1/2}\mathbf{x}_i^T \mathbf{t}))\right.\right. \\ \left.\left. + \frac{f(0)}{n} \mathbf{t}^T \mathbf{C}_n \mathbf{t}; ||\mathbf{t}|| \leq D\right\} \geq A_\eta n^{-v}\right) < n^{-\eta} \end{aligned} \quad (34)$$

and

$$\begin{aligned} P\left(\sup\left\{\left|\sum_{i=1}^n x_{ij}(\psi_{L_1}(E_i - n^{-1/2}\mathbf{x}_i^T \mathbf{t}) - \psi_{L_1}(E_i) \right.\right. \\ \left.\left. + 2f(0)n^{-1/2}\mathbf{x}_i^T \mathbf{t})\right|; ||\mathbf{t}|| \leq D\right\} \geq A_\eta n^{-v}\right) < n^{-\eta}, j = 1, \dots, p, \end{aligned} \quad (35)$$

for some  $v > 0$  and arbitrary  $D > 0$ .

**Proof:** The proof of the first assertion is a simple modification of Lemma 1 in Gutenbrunner et al (1993) and Theorem 1 in Hušková (1994c), while the second assertion follows from Theorem 2 in Hušková (1994c).  $\square$

**Lemma 2** *Let assumptions (i) - (iv) be satisfied then for any  $\eta > 0$  there exists  $A_\eta > 0$  and  $n_\eta$  such that for all  $n \geq n_\eta$*

$$P\left(\left\|\mathbf{C}_k^{1/2}(\beta_{k,L_1} - \beta) - \frac{1}{2f(0)}\mathbf{C}_k^{-1/2}\sum_{i=1}^k \mathbf{x}_i\psi_{L_1}(E_i)\right\| \geq A_\eta k^{-v}\right) < k^{-\eta}, k \leq n, \quad (36)$$

$$P\left(\left\|\mathbf{C}_k^{*1/2}(\beta_{k,L_1}^* - \beta) - \frac{1}{2f(0)}\mathbf{C}_k^{*-1/2}\sum_{i=k+1}^n \mathbf{x}_i\psi_{L_1}(E_i)\right\| \geq A_\eta(n-k)^{-v}\right) < (n-k)^{-\eta}, k < n, \quad (37)$$

for some  $v > 0$  and arbitrary  $D > 0$ .

**Proof:** The proof follows the line of the proofs of Lemma 1 in Gutenbrunner et al (1993), Theorem 4 in Hušková (1994c) and we apply Lemma 1 of the present paper.  $\square$

**Lemma 3** *Let assumptions (i) - (iv) be satisfied then*

$$\lim_{n \rightarrow \infty} P\left(a(\log n)\left(\max_{1 \leq k < n} \left\{\left(\sum_{i=1}^k \mathbf{x}_i\psi_{L_1}(E_i) - \mathbf{C}_k\mathbf{C}_n^{-1}\sum_{i=1}^n \mathbf{x}_i\psi_{L_1}(E_i)\right)^T \mathbf{C}_k^{-1}\mathbf{C}_m\mathbf{C}_k^{*-1}\right\}^{1/2} \leq t + b_p(\log n)\right)\right) = \exp\{-2\exp\{-t\}\}, t \in R^1. \quad (38)$$

**Proof:** The proof follows the line of Theorem 1.1 in Horváth (1995).  $\square$

**Lemma 4** *Let assumptions (i), (ii) and (iv) be satisfied then, as  $n \rightarrow \infty$ ,*

$$\left\{\mathbf{C}_n^{-1/2}\sum_{i=1}^{[nt]} \mathbf{x}_i\psi_{L_1}(E_i); t \in (0, 1)\right\} \rightarrow^D \{(W_1(t), \dots, W_p(t))^T; t \in (0, 1)\}, \quad (39)$$

where  $\{W_1(t); t \in (0, 1)\}, \dots, \{W_p(t); t \in (0, 1)\}$  are independent standardized Wiener processes. If, moreover, (31) is fulfilled then

$$\lim_{n \rightarrow \infty} P\left(a(\log(n/G))\left\{\max_{1 \leq k \leq n-G} \left|\sum_{i=k+1}^{k+G} \psi_{L_1}(E_i)\right|\right\} \leq t + b_1(\log(n/G)) + \log 2\right) \quad (40)$$

$$= \exp\{-2 \exp\{-t\}\}, t \in R^1$$

and

$$\begin{aligned} \lim_{n \rightarrow \infty} P\left(a(\log(n/G)) \left\{ \max_{G < k \leq n-G} \left| \sum_{i=k+1}^{k+G} \psi_{L_1}(E_i) - \sum_{i=k-G+1}^k \psi_{L_1}(E_i) \right| \right\} \right. \\ \left. \leq t + b_1(\log(n/G) + \log 3) \right) = \exp\{-2 \exp\{-t\}\}, t \in R^1. \end{aligned} \quad (41)$$

**Proof:** Since  $\sum_{i=1}^{[nt]} \mathbf{x}_i \psi_{L_1}(E_i)$  is the vector of sums of independent random variables with zero mean and finite third absolute moment and since (i), (ii) and (iv) are fulfilled the assertion (39) can be derived using standard arguments. The assertions (40) and (41) are proved, e.g., in Chen (1988).  $\square$

**Proof of Theorem 1:** We sketch the proof for  $T_{n,L_1}^{(1)}$  only, the proof for  $T_{n,L_1}^{(j)}, j = 2, 3$ , is omitted because it follows the same line.

Since Lemma 3 it suffices to show that  $T_{n,L_1}^{(1)}$  has the same limit distribution as

$$\max_{1 \leq k < n} L_{n,k}$$

where

$$\begin{aligned} L_{n,k} = \left\{ \left( \sum_{i=1}^k \mathbf{x}_i \psi_{L_1}(E_i) - \mathbf{C}_k \mathbf{C}_n^{-1} \sum_{i=1}^n \mathbf{x}_i \psi_{L_1}(E_i) \right)^T \mathbf{C}_k^{-1} \mathbf{C}_m \mathbf{C}_k^{*-1} \right. \\ \left. \left( \sum_{i=1}^k \mathbf{x}_i \psi_{L_1}(E_i) - \mathbf{C}_k \mathbf{C}_n^{-1} \sum_{i=1}^n \mathbf{x}_i \psi_{L_1}(E_i) \right) \right\}^{1/2}. \end{aligned}$$

Put

$$\begin{aligned} V_{n,k} = \left\{ 2\hat{f}(0) \left( - \sum_{i=1}^k |Y_i - x_i^T \beta_{k,L_1}| \right. \right. \\ \left. \left. - \sum_{i=k+1}^n |Y_i - x_i^T \beta_{k,L_1}^*| + \sum_{i=1}^n |Y_i - x_i^T \beta_{n,L_1}| \right) \right\}^{1/2}, k = 1, \dots, n-1. \end{aligned}$$

Applying Lemma 1 and Lemma 2 we get after tedious but straightforward calculations that, as  $n \rightarrow \infty$ ,

$$\max_{(\log n)^a \leq k \leq n - (\log n)^a} |L_{n,k} - V_{n,k}| = o_p((\log \log n)^{-1/2}) \quad (42)$$

for all  $a > 0$ . Moreover, using standard tools we receive also that

$$\max_{1 \leq k < (\log n)^a} (L_{n,k} + V_{n,k}) = o_p(\sqrt{\log \log n}) \quad (43)$$



and

$$\max_{n-(\log n)^a \leq k < n} (L_{n,k} + V_{n,k}) = o_p(\sqrt{\log \log n}) \quad (44)$$

for some  $a > 0$ .

Combining ((42) - (44) we get that the limit distribution of  $T_{n,k}^{(1)}$  is the same as  $\max_{1 \leq k < n} L_{n,k}$ . The assertion (26) follows.  $\square$

**Proof of Theorem 2:** Using Lemma 1 and Lemma 2 we receive that, as  $n \rightarrow \infty$ ,

$$\max_{1 \leq k < n} \left\| C_n^{-1/2} (\mathbf{S}_{k,L_1} - \sum_{i=1}^k \mathbf{x}_i \psi_{L_1}(E_i)) \right\| = o_p(1)$$

and

$$\max_{1 \leq k < n} \left| n^{-1/2} (S_{k,L_1}^o - \sum_{i=1}^k \psi_{L_1}(E_i)) \right| = o_p(1).$$

The proof can be then finished using classical theorems on the weak convergence of functionals of partial sums of independent random variables.  $\square$

**Proof of Theorem 3** By Lemma 1 and Lemma 2 we get after some standard steps that, as  $n \rightarrow \infty$ ,

$$\max_{1 \leq k < n-G} \left| n^{-1/2} (S_{k+G,L_1}^o - S_{k,L_1}^o - \sum_{i=k+1}^{k+G} \psi_{L_1}(E_i)) \right| = o_p((\log n)^{-1/2}).$$

The assertions (32) and (33) then follows from (40) and (41) in Lemma 4.  $\square$

## References

- [1] Antoch, J. and Hušková, M. (1992). Change point problem. In *Computational Aspects of Model Choice*, Ed. J. Antoch, pp. 11-38. Heidelberg: Physica Verlag.
- [2] Broemling, L.D. and Tsurumi, H. (1987). *Econometrics and Structural change over time* (with discussion). Marcel Dekker.
- [3] Brown, R.L., Durbin, J. and Evans, J.M. (1975). Techniques for testing the constancy of regression relationships over time (with discussion). *J. R. Statist. Soc. B* **37** 149–192.
- [4] Chen, X. (1988). Inference in a simple change-point problem. *Scientia Sinica A* **31**, 654–667.
- [5] Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change Point Analysis*. New York: Wiley.

- [6] Gombay, E. and Horváth, L. (1994). Limit theorems for change in linear regression. *J. Mult. Anal.* **48**, 43–69.
- [7] Gutenbrunner, C., Jurečková, J., Koenker, R. and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *J. Nonpar. Statist.* **2** 307–331.
- [8] Hackl, P. (1980). *Testing the Constancy of Regression Relationships over Time*. Göttingen: Vandenhoeck and Ruprecht.
- [9] Hackl, P. Ed. (1989). *Statistical Analysis and Forecasting of Economic Structural Change*. IIASA, Springer Verlag.
- [10] Hackl, P. and Westlund, A.H. Eds. (1991). *Economic Structural Change; Analysis and Forecasting*. IIASA, Springer Verlag.
- [11] Horváth, L. (1993). The maximum likelihood method for testing changes in parameters of normal observations. *Ann. Statist.* **21** 671–680.
- [12] Horváth, L. (1995). Detecting changes in linear regression. *Statistics* **26** 189–208.
- [13] Horváth, L. (1993). Limit theorems for the union intersection test. *J. Statist. Plan. Infer.* **44** 133–148.
- [14] Hušková, M. (1990a). Some asymptotic results for robust procedures for testing the constancy of regression models over time. *Kybernetika* **26** 392–403.
- [15] Hušková, M. (1990b). Asymptotics for robust MOSUM. *Commentationes Mathematicae Universitatis Carolinæ* **31** 345–356.
- [16] Hušková, M. (1994a). Nonrecursive procedures for detecting change in simple linear regression models. In *New Progress in Probability and Statistics*, Eds. M.L. Puri and J.P. Vilaplana, pp. 105–126.
- [17] Hušková, M. (1994b). Nonparametric procedures for a change in simple linear regression models. *J. Applied Statistical Sciences* **1** 337–358.
- [18] Hušková, M. (1994c). Sequential procedures based on rank regression scores. *J. Nonpar. Statist.* **3** 337–358.
- [19] James, B., James, K.L. and Siegmund, D. (1987). Tests for change-point. *Biometrika* **74** 71–83.
- [20] Jandhyala, V.K. and MacNeill, I.B. (1989). Residual partial sum limit process for regression models with applications to detecting parameter changes at unknown times. *Stochastic Processes and their Applications* **33** 309–323.
- [21] Jandhyala, V.K. and MacNeill, I.B. (1991). Tests for parameter changes at unknown times in linear regression models. *J. Statist. Plann. Infer.* **27** 291–316.
- [22] Jandhyala, V.K. and MacNeill, I.B. (1992). On testing for the constancy of regression coefficients under random walk and change-point

- alternatives. *Econometric Theory* **8** 501–517.
- [23] Kim, Hyune-Ju and Siegmund, D. (1989). The likelihood ratio test for a change point in simple linear regression. *Biometrika* **76** 409–423.
  - [24] Maddala, G.S. (1977). *Econometrics*. New York: McGraw-Hill.
  - [25] Ploberger, W. and Krämer, W. (1992). The CUSUM test with OLS residuals. *Econometrica* **60** 271–285.
  - [26] Quandt, R.E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Am. Statist. Assoc.* **53** 873–880.
  - [27] Quandt, R.E. (1960). Tests of the hypothesis that a linear regression system obey two separate regimes. *J. Am. Statist. Assoc.* **55** 342–330.
  - [28] Sen, P.K. (1980). Asymptotic theory of some tests for a possible change in the regression slope occurring at an unknown time point. *Z. Wahr. Verw. Geb.* **52** 203–218.
  - [29] Sen, P.K. (1981). *Sequential Nonparametrics*. New York: Wiley.
  - [30] Sen, P.K. (1982). Asymptotic theory of some tests for constancy of regression relationships over time. *Statistics* **13** 21–31.
  - [31] Sen, P.K. (1984). Recursive M-tests for the constancy of multivariate regression relationships over time. *Sequential Analysis* **3** 191–211.
  - [32] Worsley, K.J. (1983). Testing for a two-phase multiple regression. *Technometrics* **25** 35–42.

# Inadmissibility of robust estimators with respect to $L_1$ norm

**Jana Jurečková**

*Charles University, Prague, Czech Republic*

**L.B. Klebanov**

*Institute of Mathematical Geology, St. Petersburg, Russia*

*Abstract:* We show that robust M-estimators as well as equivariant estimators which do not depend on the extreme observations are inadmissible estimators of the location with respect to the  $L_1$  loss function for a broad class of distributions. As a consequence, it implies that the sample median is inadmissible as an estimator of the location of the double-exponential distribution.

*Key words:* Admissibility, equivariant estimator, Pitman estimator,  $M$ -estimator,  $L_1$  norm.

AMS subject classification: 62C15, 62F11, 62F35.

## 1 Introduction

Let  $X_1, \dots, X_n$  be a random sample from a distribution with the absolutely continuous distribution function  $F(x - \theta)$ ,  $\theta \in \mathbb{R}^1$ . The problem is that of estimating the parameter  $\theta$ . We shall assume that the loss  $L(t, \theta)$  incurred when estimating  $\theta$  by  $t$  depends only on  $|t - \theta|$ , *i.e.*

$$L(t, \theta) = L(|t - \theta|). \quad (1)$$

Then it is natural to restrict considerations to the estimators equivariant with respect to the shift in location, *i.e.* satisfying

$$T_n(X_1 + c, \dots, X_n + c) = T_n(X_1, \dots, X_n) + c \quad \forall c \in \mathbb{R}^1 \quad \text{and} \quad \forall \mathbf{X} \in \mathbb{R}^n. \quad (2)$$

Let  $\mathcal{T}$  denote the family of all equivariant estimators.

Different measures of performance of equivariant estimators were investigated. Among them, the probability

$$P_\theta(|T_n - \theta| > a) \quad (3)$$

that the absolute error exceeds  $a > 0$ , was considered by several authors, either for  $n \rightarrow \infty$  and  $a > 0$  fixed or  $n$  fixed and  $a \rightarrow \infty$ . Bahadur [2], [3], Fu [4] and Sievers [12] studied the limit

$$\lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log P_\theta(|T_n - \theta| > a) \right\} = e, \quad a \text{ fixed} \quad (4)$$

as a measure of performance of  $T_n$ . Sievers [12], who calculated the limits  $e$  for several estimators and several distribution shapes, found the sample median less efficient than the sample mean not only for normal but also for logistic distribution and even for the double exponential distribution for sufficiently large  $a$ . Similar phenomenon was observed by Jurečková [5] who considered the measure of performance

$$B(a; T_n) = \frac{-\log P_0(|T_n| > a)}{-\log P_0(|X_1| > a)} \quad (5)$$

under  $n$  fixed and  $a \rightarrow \infty$ . The estimators which trim-off the extreme observations were found more robust but less efficient than  $\bar{X}_n$  for densities with exponential tails, including the double exponential. Akahira and Takeuchi [1] computed the loss of information caused by trimming the extreme order statistics in the double exponential population.

Denote

$$\mathbf{Y} = (Y_1, \dots, Y_n) \text{ where } Y_i = X_i - X_1, \quad i = 1, \dots, n \quad (6)$$

the maximal invariant of  $\mathbf{X}$  with respect to the group of translations of  $X_1, \dots, X_n$ . Then the minimum risk equivariant estimator  $T_n^*$  (*Pitman estimator*, MRE) exists provided there exists at least one  $T_n \in \mathcal{T}$  with finite risk; then  $T_n^*$  could be written in the form

$$T_n^*(\mathbf{X}) = T_n(\mathbf{X}) - v^*(\mathbf{Y}) \quad (7)$$

where  $v^*(\mathbf{Y})$  satisfies

$$\mathbb{E}_0 L(|T_n(\mathbf{X}) - v^*(\mathbf{Y})|) = \min_{v(\mathbf{Y})} \mathbb{E}_0 L(|T_n(\mathbf{X}) - v(\mathbf{Y})|) \quad (8)$$

with the minimum taken over all functions  $v(\mathbf{y})$ .

If the loss function is quadratic,  $L_2(t, \theta) = (t - \theta)^2$ , the minimum risk estimator has the form

$$T_n^{(2)}(\mathbf{X}) = T_n(\mathbf{X}) - \mathbb{E}_0(T_n(\mathbf{X})|\mathbf{Y})$$

and it equals to the sample mean  $\bar{X}$  provided  $F$  is normal. Conversely, Kagan, Linnik and Rao [9] proved that, provided  $\bar{X}_n$ ,  $n \geq 3$ , is the minimum risk estimator of  $\theta$  for some  $F$  with respect to the quadratic loss, then  $F$  is normal. Otherwise speaking, the equality  $\mathbb{E}_0(\bar{X}|\mathbf{Y}) = 0$  characterizes the normal distribution.

If  $f$  is not normal, the minimum risk estimator of  $\theta$  is typically nonlinear and only the sample mean is a sum of independent summands. However, many estimators admit asymptotic representations of the type

$$T_n(\mathbf{X}) = \theta + \frac{1}{n} \sum_{i=1}^n \psi(X_i - \theta) + o_p(n^{-1/2}) \quad \text{as } n \rightarrow \infty \quad (9)$$

with appropriate functions  $\psi$ . Especially, if  $T_n(\mathbf{X})$  is an asymptotically efficient estimator, then  $\psi(x) = -\frac{1}{I(f)} \frac{f'(x)}{f(x)}$ , with  $f$  being the density of  $F$  and  $I(f)$  its Fisher information. A systematic study of the representations of this type could be found in [7]. Jurečková and Milhaud [6] recently proved that if the equality  $\mathbb{E}_0(\sum_{i=1}^n \psi(X_i)|\mathbf{Y}) = 0$  holds for  $n \geq 4$  and for a function  $\psi$  satisfying some regularity conditions, then  $\psi(x) = c \frac{f'(x)}{f(x)}$ ,  $x \in \mathbb{R}^1$ , where  $c$  is a constant and  $f$  is the density of  $F$ . This result indicates that not only asymptotically, but also in the finite sample case, many properties of  $\frac{1}{nI_f} \sum_{i=1}^n \left(-\frac{f'(X_i)}{f(X_i)}\right)$  under  $F$  are in correspondence with the respective properties of the sample mean under the normal distribution.

In the present paper, we shall consider the performance of some robust estimators with respect to the  $L_1$  loss, *i.e.*,

$$L_1(t, \theta) = |t - \theta|. \quad (10)$$

Zinger, Kagan and Klebanov [13] and Kagan and Zinger [10] (see also [9], Section 7.9) proved that if  $\bar{X}_n$  is the minimum risk estimator of  $\theta$  with respect to  $L_1$ -loss for  $f(x - \theta)$ ,  $f$  unimodal and  $n \geq 6$ , then the underlying distribution is normal. This means that, for the normal distribution,  $\bar{X}$  is the minimum risk estimator of  $\mathcal{T}$  with respect to both  $L_1$  and  $L_2$  loss functions and in both cases its admissibility it is a characteristic property of the normal distribution.

If  $f$  is unknown then we prefer robust estimators which are not connected with a fixed density shape. However, if we know  $f$ , then we are interested in admissible estimators whose risk cannot be uniformly improved. We shall show that the robust estimators are not admissible under the  $L_1$  norm for a broad class of densities.

## 2 Inadmissibility of trimmed estimators

Let  $X_1, \dots, X_n$  be a random sample from a distribution with the continuous density  $f(x - \theta)$  such that  $f(x) > 0$ ,  $x \in \mathbb{R}^1$ . Let  $X_{n:1} \leq \dots \leq X_{n:n}$  be the order statistics corresponding to  $X_1, \dots, X_n$ . Let  $\mathcal{T}^* \subset \mathcal{T}$  denote the set of equivariant estimators satisfying the following condition:

(A1) If  $T_n = T_n(X_{n:i_1}, \dots, X_{n:i_k})$ ,  $i_1 < \dots < i_k$ ,  $1 \leq k \leq n$ , then  $X_{n:i_1} \leq T_n \leq X_{n:i_k}$ .

(A2)  $T_n(0, \dots, 0) = 0$ .

The following theorem shows that the trimmed estimators as well as the M-estimators with a score function constant outside a bounded interval are inadmissible for unimodal densities.

**Theorem 1** *Let  $X_1, \dots, X_n$ ,  $n \geq 5$ , be independent observations from a distribution with the density  $f(x - \theta)$ , where  $f(x) > 0$ ,  $x \in \mathbb{R}^1$ , is unimodal, i.e. increasing for  $x < 0$  and decreasing for  $x > 0$ .*

(i) *Let  $T_n \in \mathcal{T}^*$  be an equivariant estimator,  $T_n(X_1, \dots, X_n)$ , continuous in each argument, constant with respect to  $X_{n:1}, X_{n:2}, X_{n:n-1}$  and  $X_{n:n}$ , but uniquely determined as a function of  $X_{n:3}, \dots, X_{n:n-2}$ . Then  $T_n$  is inadmissible estimator of  $\theta$  with respect to the loss (10).*

(ii) *Let  $M_n$  be an M-estimator generated by a continuous non-decreasing function  $\psi$  such that  $\psi(x) = \psi(c_1)$  for  $x \leq c_1$  and  $\psi(x) = \psi(c_2)$  for  $x \geq c_2$ ,  $c_1 < 0 < c_2$ , as*

$$M_n = \frac{1}{2}(M_n^+ + M_n^-) \quad \text{where} \quad (11)$$

$$M_n^- = \sup\{t : \sum_{i=1}^n \psi(X_i - t) > 0\}, \quad M_n^+ = \inf\{t : \sum_{i=1}^n \psi(X_i - t) < 0\}.$$

*Then  $M_n$  is inadmissible as an estimator of  $\theta$  with respect to the loss (10).*

**Proof:** In the case of  $L_1$  norm, (7) specializes to

$$T_n^* = T_n - \text{med}_0(T_n | \mathbf{y}) \quad (12)$$

where  $\text{med}_0(T_n | \mathbf{y})$  stands for any conditional median of  $T_n$  given the maximal invariant  $\mathbf{y}$  under  $\theta = 0$ . Hence,  $T_n$  is the MRE provided

$$\text{med}_0(T_n | \mathbf{y}) = 0 \quad (13)$$

and the median is unique.

(i) Let  $T_n \in \mathcal{T}$  be uniquely determined and do not depend on  $X_{n:1}, X_{n:2}, X_{n:n-1}$  and  $X_{n:n}$ . Denote

$$\mathbf{Y} = (Y_1, \dots, Y_n) = (X_1 - T_n, \dots, X_n - T_n). \quad (14)$$

Then  $\mathbf{Y}$  is the maximal invariant for the group of translations of  $X_1, \dots, X_n$ . The conditional distribution of  $T_n$  given  $\mathbf{Y} = \mathbf{y}$  has the density

$$g(t|\mathbf{y}) = \frac{\prod_{i=1}^n f(t + y_i)}{\int_{-\infty}^{\infty} \prod_{i=1}^n f(z + y_i) dz}. \quad (15)$$

The condition (13) rewrites in the following way:

$$\int_{-\infty}^0 \prod_{i=1}^n f(t + y_i) dt = \int_0^{\infty} \prod_{i=1}^n f(t + y_i) dt \quad \text{a.s. } [F]. \quad (16)$$

In view of continuity of the density  $f$  and the estimator  $T_n$  previous equation holds not only for almost all but for all  $y_i$ .

Denoting  $w(t) = \text{sign } t$ ,  $t \in \mathbb{R}^1$ , we rewrite (16) in the form

$$\int_{-\infty}^{\infty} w(t) \prod_{i=1}^n f(t + y_i) dt = 0. \quad (17)$$

Denote  $A = \{\mathbf{y} : y_2 = \dots = y_{n-2} = 0, y_1 \leq y_2 \leq 0, y_n \geq y_{n-1} \geq 0\}$ . Then  $T_n(\mathbf{y}) = 0$  for  $\mathbf{y} \in A$  independently of the values of  $y_1, y_2, y_{n-1}, y_n$  and (17) takes on the form

$$\int_{-\infty}^{\infty} w(t) f(t + y_1) f(t + y_2) f(t + y_{n-1}) f(t + y_n) f^{n-2}(t) dt = 0, \quad \mathbf{y} \in A. \quad (18)$$

Differentiating (18) in  $y_\nu$ ,  $\nu = 1, 2, n-1, n$  gives

$$\int_{-\infty}^{\infty} w(t) f'(t + y_\nu) \prod_{i \neq \nu} f(t + y_i) f^{n-4}(t) dt = 0, \quad \nu = 1, 2, n-1, n. \quad (19)$$

Integrating the left-hand side of (19) by parts for  $\nu = 1$  and using (19) for  $\nu = 2, n-1, n$ , we obtain

$$(n-4) \int_{-\infty}^{\infty} w(t) f'(t) \prod_{i=1,2,n-1,n} \frac{f(t + y_i)}{f(y_i)} f^{n-5}(t) dt = 2f^{n-4}(0) \quad (20)$$

If we especially take three following choices of  $\mathbf{y} \in A$ :

$$\begin{aligned} y_1 &= y_2 = u < 0, y_{n-1} = y_n = 0 \\ y_1 &= u < 0, y_2 = 0, y_{n-1} = 0, y_n = v > 0 \\ y_1 &= y_2 = 0, y_{n-1} = y_n = v > 0, \end{aligned}$$



and subtract twice the second equality from the sum of the first and the third ones, we get

$$\int_{-\infty}^{\infty} w(t)f'(t) \left[ \frac{f(t+u)}{f(u)} - \frac{f(t+v)}{f(v)} \right]^2 f^{n-3}(t) dt = 0. \quad (21)$$

Because  $w(t)f'(t) > 0$  for  $t \neq 0$ , (21) implies that

$$\frac{f(t+u)}{f(u)} = \frac{f(t+v)}{f(v)} \quad (22)$$

holds for all  $t \neq 0$ ,  $u < 0$ ,  $v > 0$ , and this in turn implies that

$$\frac{f(t+u)}{f(u)} = \frac{f(t+v)}{f(v)} \quad (23)$$

holds for all  $t \neq 0$ ,  $u, v \in \mathbb{R}^1$ . By the Cauchy equation, the only function satisfying (23) is either the exponential function or the constant. By the unimodality assumption on  $f$ ,  $\forall v > 0$  there exists  $u < 0$  such that  $f(u) = f(v)$ , and then (22) leads to the constant  $f$ , what is a contradiction. Hence, there exists at least one  $\mathbf{y}^* \in A$  such that either

$$\int_{-\infty}^0 \prod_{i=1}^n f(t+y_i^*) dt < \int_0^{\infty} \prod_{i=1}^n f(t+y_i^*) dt \quad (24)$$

[or the opposite inequality] holds for  $\mathbf{y}^*$ . Then, because of the continuity, (24) [or the opposite inequality] holds in a neighborhood of  $\mathbf{y}^*$ , hence (16) is not true *a.s.*  $[F]$  and  $T_n$  is not admissible.

(ii) Let  $M_n$  be the M-estimator defined in (11). Put  $Y_i = X_i - M_n$ ,  $i = 1, \dots, n$ . Then  $\mathbf{Y}$  is the maximal invariant with respect to the group of translations and the conditional density of  $M_n$  given  $\mathbf{Y} = \mathbf{y}$  has the form (15). Analogously as in the part (i),  $M_n$  would be the MRE under the condition (17). Let  $B = \{\mathbf{y} : y_5 = \dots = y_n = 0\}$ ; proceed analogously as in (19) - (21) and take successively the following choices of  $y_1, \dots, y_4$  :

$$\begin{aligned} y_1 &= y_2 = c_1 - u, \quad y_3 = y_4 = c_1 - u', \\ y_1 &= c_1 - u, \quad y_2 = c_2 + v, \quad y_3 = c_1 - u', \quad y_4 = c_2 + v', \\ y_1 &= y_2 = c_2 + v, \quad y_3 = y_4 = c_2 + v', \end{aligned}$$

$u, u', v, v' \geq 0$ . Analogously as in the part (i), we arrive at the equation

$$\int_{-\infty}^{\infty} w(t)f'(t) \frac{f(t+c_1-u)f(t+c_1-u')}{f(c_1-u)f(c_1-u')} -$$

$$\frac{f(t+c_2+v)f(t+c_2+v')^2}{f(c_2+v)f(c_2+v')} f^{n-5}(t) dt = 0 \quad (25)$$

$\forall t \in \mathbb{R}^1, u, u', v, v' \geq 0$ . Quite analogously we get

$$\int_{-\infty}^{\infty} w(t) f'(t) \frac{f(t+c_1-u)f(t+c_2+v')}{f(c_1-u)f(c_2+v')} - \frac{f(t+c_2+v)f(t+c_1-u')^2}{f(c_2+v)f(c_1-u')} f^{n-5}(t) dt = 0 \quad (26)$$

$\forall t \in \mathbb{R}^1, u, u', v, v' \geq 0$ . Similarly as in part (i), we first conclude that the density  $f$  should be then either exponential or constant in the tails and (25) finally leads to the constant tails, what is a contradiction with the conditions imposed on  $f$ . Thus, there exists  $\mathbf{y}^* \in B$  and hence also its neighborhood satisfying either (24) or the opposite inequality; we conclude that  $M_n$  cannot be an admissible estimator of  $\theta$  with respect to  $L_1$  loss.  $\square$

Notice that Theorem 1 covers the trimmed L-estimators, the sample median as well the linear combinations of several (non-extreme) sample quantiles; it also covers the Huber estimator and the related M-estimators. The results also imply that the sample median is not admissible even for the double exponential distribution for which it is the maximum likelihood estimator. Similarly, while Huber's M-estimator is MLE for the contaminated normal distribution, it is not admissible for the same.

## Acknowledgements

For Jana Jurečková, the research was supported by the Charles University Grant GAUK 198/1996, and for L. B. Klebanov, the research was partially supported by RFFI Grants 95-01-01260, 96-01-00852.

## References

- [1] Akahira, M. and Takeuchi, K. (1995). *Non-Regular Statistical Estimation*. Lecture Notes in Statistics 107, Springer-Verlag.
- [2] Bahadur, R.R. (1960). Asymptotic efficiency of tests and estimates. *Sankhyā* **22** 229-252.
- [3] Bahadur, R.R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303-324.
- [4] Fu, J.C. (1975). The rate of convergence of consistent point estimators. *Ann. Statist.* **3** 234-240.

- [5] Jurečková, J. (1981). Tail-behavior of location estimators. *Ann. Statist.* **9** 578-586.
- [6] Jurečková, J. and Milhaud, X. (1996). Characterization of distributions in invariant models. Submitted.
- [7] Jurečková, J. and Sen, P.K. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations*. New York: Wiley.
- [8] Kagan, A.M., Linnik, Yu.V. and Rao, C.R. (1965). On a characterization of the normal law based on a property of the sample average. *Sankhyā A* **27** 405-406.
- [9] Kagan, A.M., Linnik, Yu.V. and Rao, C.R. (1972). *Characterization Problems of Mathematical Statistics* (in Russian). Moscow: Nauka
- [10] Kagan, A.M. and Zinger, A.A (1971). Sample mean as an estimator of location parameter. *Sankhyā A* **33** 351-358.
- [11] Lehmann, E.L. (1983). *Theory of Point Estimation*. New York: Wiley.
- [12] Sievers, G.L. (1978). Estimation of location; a large deviation comparison. *Ann. Statist.* **6** 610-618.
- [13] Zinger, A.A., Kagan, A.M. and Klebanov, L.B. (1969). Sample mean as an estimator of the shift parameter with non-quadratic loss functions (in Russian). *Doklady AN SSSR* 186, 1.

# Properties of $L^1$ residuals

Stephan Morgenthaler

*Swiss Federal Institute of Technology, Lausanne, Switzerland*

*Abstract:* The paper discusses the behavior of residuals from least-absolute-deviations (or  $L^1$ ) fits of linear models. Particular emphasis is given to data arising by way of designed experiments. The paper argues that the  $L^1$  method of fitting such models should be discouraged. The method is inefficient when compared to other robust methods while not being any simpler to compute. The residuals obtained by  $L^1$  fitting exhibit several weaknesses. First of all they are ambiguous in the sense that there are a multitude of  $L^1$  fits, sometimes quite far apart. Second, typical algorithms produce as many exact zero residuals as there are contrasts fitted. As a result, the non zero residuals do not give an accurate reflection of the errors that occurred during the experimental runs.

*Key words:* Least absolute deviations, factorial designs, outliers detection, uniqueness of fit.

AMS subject classification: 62K15, 62J05.

## 1 Introduction

Let  $y = X\theta + \varepsilon$  be a linear model with uncorrelated, centered, and homoskedastic errors  $\varepsilon_1, \dots, \varepsilon_n$ . As indicated, we take  $n$  to be the number of observations, whereas  $p$  denotes the dimension of the regression parameter  $\theta$ . The least-squares residuals are  $r = (I - H)y$ , where  $\hat{y} = Hy = X(X^T X)^{-1}X^T y$  is the least-squares fit. If the error distribution has two moments, it follows that  $E(r) = X\theta - HX\theta = 0$ , and  $Var(r) = \sigma^2(I - H)$ . The use of such residuals for outlier detection and other diagnostic purposes has been explored in great detail in the statistical literature (see for example Belsley, Kuh and Welsch, 1980; Cook and Weisberg, 1982).

Residuals from an  $L^1$  fit are not so easily described. Throughout this article, we denote by  $\tilde{\theta}$  a parameter fit obtained by minimizing the least-

absolute-deviations and designate by  $e$  the corresponding residuals  $e = y - X\tilde{\theta}$ . Because  $\mathbb{R}^\times$  equipped with the  $L^1$  norm is only a weakly convex normed linear space, the best approximation to  $y$  of the form  $X\tilde{\theta}$  is in general not unique. In fact, the best approximations form a convex set in the  $p$ -dimensional column space of  $X$ . It is widely-known that among the best approximations there is always one, for which at least  $p$  of the components of  $e$  are exactly equal to zero. The algorithms based on linear programming techniques always identify one of these solutions, because they correspond to extremal points of the linear programming problem.

For Gaussian errors, the least-squares fit  $\hat{\theta}$  is fully efficient, whereas the least-absolute-deviation fit  $\tilde{\theta}$  reaches an asymptotic efficiency of  $2/\pi = 63.7\%$ . In balanced factorial models the element  $h_{ij}$  of  $H$ , proportional to the covariance of the least-squares fit  $\hat{y}_i$  and  $\hat{y}_j$ , depends in a simple way on the factor settings at runs  $i$  and  $j$ . In the case of a two-way ANOVA with factors  $f_1$  and  $f_2$ , for example, there are four cases, distinguished by the comparison of  $(f_{1i}, f_{2i})$  and  $(f_{1j}, f_{2j})$ . In particular, the diagonal elements  $h_{ii}$  are all equal to  $p/n$ , where  $p$  is the dimension of the column space of  $X$ . In the following, we restrict our discussion to this case of a balanced factorial model.

The asymptotic behavior of the residuals is for  $L^1$  and  $L^2$  the same, as long as  $p/n$  tends to zero with increasing  $n$ . The residuals are asymptotically equivalent to a sample from the error distribution. Asymptotic considerations are, however, of minor interest when one discusses properties of residuals. The case  $p \approx n$  is of much more practical concern.

## 2 Identifying a small flock of outliers

Least-squares residuals have a tendency to behave much like a sample from a Gaussian distribution. Stem-and-leave plots or normal plots do often not reveal anything of interest. This is due to the dependence imposed on the residuals by the requirement that  $r^T X = 0$ . For that reason, glaring error structures will be lost or not faithfully translated into residual structures. An example of this sort concerns the presence of a few **outliers** among the measurement errors. To illustrate what happens, suppose the residual  $r$  has a fixed size, e.g.  $r^T r = 1$  and we seek to maximize  $w^T r$  for a fixed vector of component weights  $w$ . The solution to this constrained optimisation problem yields a maximal value of  $w^T w - w^T H w$  achieved, when  $r \propto (I - H)w$ . Thus, if we maximize a single component of a (unit) least-squares residual  $r$ , the largest possible value is  $\sqrt{1 - h_{ii}} = \sqrt{(n - p)/n}$ . If we maximize the sum of two components, the largest value for the sum of the  $i$ th and  $j$ th residuals is  $\sqrt{2 - h_{ii} - h_{jj} - 2h_{ij}}$ , etc.

**Example 1** *A simple illustration of these facts can be given by using the  $2^2$  main effects model. If all observations are zero except one, which is equal to  $c$ , the residuals are equal to  $\pm c/4$ . This residual vector has  $L^2$ -norm equal to  $c^2/4$  and  $c = 2$  normalizes it. The largest possible component of a unit residual is, therefore, equal to  $1/2 = \sqrt{(4-3)}/4$ , which means that the largest percentage of the  $L^2$ -norm of a residual that resides on a single component is 25%.*

In a  $3 \times 3$  main effects model, the corresponding number is  $2/3 = \sqrt{(9-5)}/9$ , which implies that the largest percentage of the  $L^2$ -norm residing in a single residuals is equal to 44.4%. These examples illustrate the fact that the ability of a design to show a single outlier by way of a large individual residual depends on the ratio  $p/n$ .

The general formula given above can be used to judge the ability of a given design to point out in a single experiment two outliers by two large residuals. It is immediately clear that this capacity is dependent on the positions (on the indices) of the outlying observations, since  $h_{ij}$  depends on  $i$  and on  $j$ .

The picture is maybe clarified, if we pose the question differently. Given a residual vector  $r$ , what maximal percentage of its  $L^2$ -norm  $r^T r$  can be explained by 1, or 2, or 3, etc. components. Let  $I = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$  denote a set of  $m$  indices. It turns out that the answer to the above query is equal to  $1 - \lambda_{\min}(I)$ , where  $\lambda_{\min}(I)$  denotes the smallest eigenvalue of the minor  $H_I$  of the hat matrix determined by the intersection of the rows and columns from  $I$ . This is easy to show and we leave it to the reader to check the statement.

**Example 2** *When the number  $m$  of residuals we wish to check is equal to 1, the minor  $H_I$  is equal to the scalar  $h_{i_1, i_1} = p/n$  and  $1 - \lambda_{\min}(I) = 1 - p/n$ , which is a result we already knew. If we pass to  $m = 2$  for the  $2^2$  design, all minors of dimension 2 have a minimal eigenvalue of  $1/2$ . The largest percentage of the total norm explained by two components, i.e., by half the components, is equal to 50%. The situation in the  $3 \times 3$  case is different. The smallest eigenvalue of 2-dimensional minors is either  $1/3$  or  $4/9$ , depending on the position of the pair within the 3 by 3 table. The maximal percentage of the total norm that can be explained by 2 of the 9 residuals is, therefore, 66.7%. Since with a single component, one can at most explain a percentage of 44.4%, this is a bit disappointing. Evidently, two outliers will result in two residuals that stick out much less than was the case with a single outlier. This kind of behavior is typical for least-squares fits.*

What is the answer to the same question in the case of least-absolute-

deviations residuals  $e$ ? For values of  $m$  smaller than the **exact-fit-point** of the  $L^1$  method, the residuals can be completely concentrated on any of set of  $m$  components. The exact-fit-point  $N_{\text{ef}}$  of an equivariant fitter can be defined as the largest number of non zero observations that one can add in any position to the vector of observations  $y = 0$  without changing the fit from  $\tilde{y} = 0$ . In the best of situations, this point is equal to or close to  $n/4 - 1$  for the least-absolute-deviations regression method (for details, see Ellis and Morgenthaler, 1992).

A thorough discussion of the break down and outlier resistance problem in designed experiments is given in Müller (1995). If we wish to be able to fit all contrasts in a given model, the crucial quantity is the maximal number of experimental runs that by themselves are not enough to determine a fit of the model. In the  $3 \times 3$  main effects design this number is equal to six, which is bigger than the five dimensions of the parameter space. Any equivariant fitter breaks down, as soon as a majority of the  $3 = 9 - 6$  remaining observations are faulty.

For  $m > N_{\text{ef}}$ , depending on the position of the outliers, different outcomes are possible.

**Example 3** In a  $3 \times 3$  design,  $N_{\text{ef}} = 1$ . For  $m = 2$  and  $m = 3$ , the following tables show some of the possibilities.

|     |     |     |
|-----|-----|-----|
| $a$ | $b$ | $0$ |
| $0$ | $0$ | $0$ |
| $0$ | $0$ | $0$ |

|     |     |     |
|-----|-----|-----|
| $a$ | $0$ | $0$ |
| $0$ | $b$ | $0$ |
| $0$ | $0$ | $0$ |

|     |     |     |
|-----|-----|-----|
| $a$ | $0$ | $0$ |
| $0$ | $b$ | $0$ |
| $0$ | $0$ | $c$ |

In the left-most case, the residual  $e$  has in general two non zero components. They can be in the same position as the outliers  $a$  and  $b$  – – this happens when they are of opposite sign – – or they can be spread to other positions, one at the third position of the first line, the other indicating the more important outlier among  $a$  and  $b$ . In the middle case, the non zero residuals are confined to the two positions where  $a$  and  $b$  are observed as long as they have the same sign. Otherwise, the  $L^1$  fit is not unique and a second and third non zero residual can pop up in the first and the last line. If  $a$  is  $-300$  and  $b$  is  $290$ , the residual table can be as disparate as the following two examples:

|        |       |     |
|--------|-------|-----|
| $-300$ | $0$   | $0$ |
| $0$    | $290$ | $0$ |
| $0$    | $0$   | $0$ |

|       |        |       |
|-------|--------|-------|
| $-10$ | $0$    | $290$ |
| $0$   | $0$    | $0$   |
| $0$   | $-290$ | $0$   |

Since the vector

|   |    |   |
|---|----|---|
| 1 | 0  | 1 |
| 0 | -1 | 0 |
| 0 | -1 | 0 |

lies in the column space of the design matrix  $X$ , one can smoothly transform between these two residual vectors without changing the  $L^1$  norm. In the right-most case, the situation is even more complex. If the outliers  $a$ ,  $b$  and  $c$  are of equal sign, the residual matrix faithfully reflects this structure. If they are of unequal signs, surprising things can happen. The observed table

|     |     |      |
|-----|-----|------|
| 300 | 0   | 0    |
| 0   | 290 | 0    |
| 0   | 0   | -250 |

leads to a unique  $L^1$  fit with residual table equal to

|      |      |   |
|------|------|---|
| 50   | -250 | 0 |
| -250 | 40   | 0 |
| 0    | 0    | 0 |

In this case, there exists an additive fit explaining the data with (merely) two sizable residuals. When  $m = 3$ , we are beyond the range, where we can generally expect to distinguish outliers from additive structure in  $3 \times 3$  tables. Most robust procedures prefer the fit found by the  $L^1$  method over the fit  $\tilde{y} = 0$ . But one can, of course, imagine procedures that are able to identify any additive structure as long as it is exactly adhered to by a majority (LMS, Rousseeuw and Leroy, 1987). Such a procedure could not distinguish between the two fits which both have at least 5 of the 9 residuals equal to zero.

### 3 Maximal residuals

The size of the largest residual will often be taken as an indication, whether faulty runs occurred during a designed experiment. As we saw in the last section, when only a small number – less than the exact-fit-point of the runs are faulty, and stick out very clearly, then the  $L^1$  fit will produce residuals that can safely be used to identify the faulty runs. Beyond this number of grossly wrong observations and when the outlyingness is less clear cut, the  $L^1$  method is not successful. We also noted in the last section that the  $L^1$  fit has two drawbacks. Firstly, **it does in general not produce a unique answer**. This may at first sight seem not to be a concern, but in the case of designed experiments, multiplicity of possible answers is very



common and the set of  $L^1$  solutions can be very varied. Secondly, the solutions found with the known algorithms produce at least  $p$  exact zeros among the residuals. Such least-absolute-deviation fits are typically not very appealing when one analyzes them in detail. The greed for exact zeros tends to make the non zero residuals “too” big.

There are several simple arguments which allow us to estimate the average inflation factor that we should expect when passing from least-squares residuals to least-absolute-deviations residuals. First, the  $L^1$  criterion will be about the same between the two solutions, i.e.  $\sum_{i=1}^n |e_i| \leq \sum_{i=1}^n |r_i|$  with rough equality for large values of  $n/p$ . If this were so and if we further imagine that the exact zeroes are created by randomly selecting the cells, then the fact that the first sum contains  $p$  exact zeros makes  $|e_i|$  on average  $n/(n-p)$  times larger than  $|r_i|$ . This is equivalent to imagining that the non-zero  $L^1$  residuals are constructed from a  $L^2$  residual to which  $p/(n-p)$  parts of  $p/n$   $L^2$  residuals are added. Both of these arguments over-estimate the inflation factor. A final check can also be made on the level of the variance. Suppose that a random selection of  $(n-p)/n$  of the  $L^2$  residuals were multiplied by the above inflation factor, i.e.,  $n/(n-p)$ , whereas the others were put equal to zero. Under such a process, the variance of the non zero  $L^1$  residuals would be equal to  $(n/(n-p))^2$  times the variance of the  $L^2$  residuals. If we take the rough equality of the  $L^2$  criterion as a guide, we would expect the variation in  $e$  to be the same as the variation in  $r$ . Since the components in  $e$  contain a mixture of  $p/n$  exact zeros with zero variance and  $(n-p)/n$  non zeros, the variation of the non zero residuals is expected to be  $n/(n-p)$  times bigger than the variation in  $r$ . This leads to an inflation factor of  $\sqrt{(n-p)/n}$ , but this time one underestimates the true size. Both factors tend to 1, as  $p/n$  tends to zero and both are true some of the time. Typically, when we have only a few degrees of freedom for the error, then the factor  $n/(n-p)$  is correct. This is the case for example, when we fit a  $2^k$  design up to the  $(k-1)$ -factor interactions. But, it is also roughly true for a  $2 \times k$  factorial design, where the number of degrees of freedom for the error is  $k/2$  and thus arbitrarily large. The last example shows, that it is also the design itself that has an influence on the behavior of the  $L^1$  method.

In the  $3 \times 3$  design, the inflation factor for the size of residuals is between  $\sqrt{9/4} = 1.5$  and  $9/4 = 2.25$ . Figure 1 illustrates what really happens for four simple designs. Note that an innocent interpretation of the  $L^1$  residuals  $e$  would quite often lead to the conclusion that outlying experimental runs were present, because of the large maximal residual size. The simple minded adjustment given above works reasonably well.

Figure 2 repeats the experiment explained in Figure 1, but this time

with errors from a contaminated Gaussian with  $1/n$  contamination having a five-fold standard deviation.

Among the 300 simulated experiments, roughly 60% resulted in the correct identification of the faulty run in the sense that the largest sized residual was indeed associated with the contaminated run. This is true for both methods of fitting. However, in the least-squares case, the largest residual does not stick out clearly when compared to the next largest one.

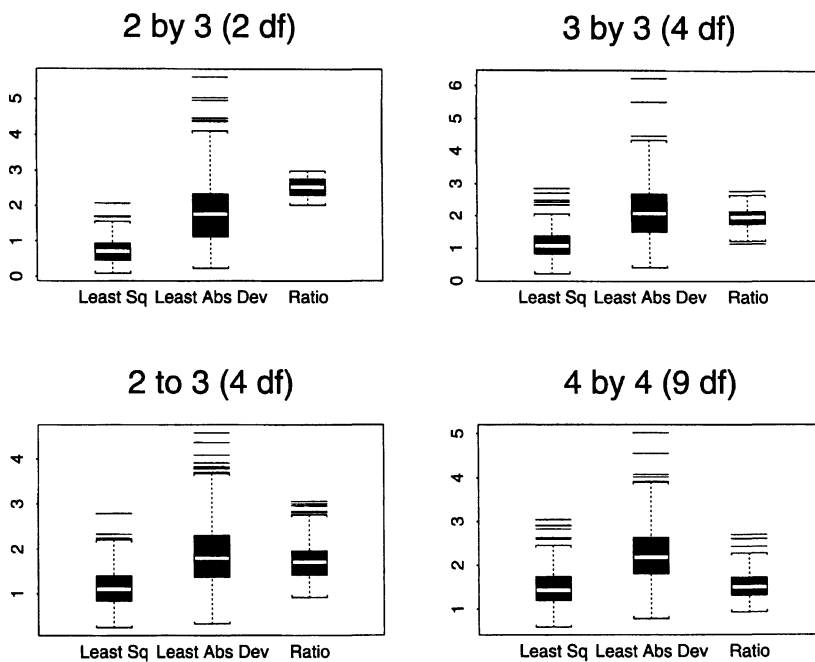


Figure 1: The figure shows the behavior of the maximal residual size in four different design. The expected inflation factors are  $(1.73, 3.00)$  for the  $2 \times 3$ ,  $(1.50, 2.25)$  for the  $3 \times 3$ ,  $(1.41, 2.00)$  for the  $2^3$  main effects and  $(1.33, 1.78)$  for the  $4 \times 4$ . The average inflation factor of the maximal residual observed in 300 replications are 2.54, 1.93, 1.61 and 1.51. The ratios between the maximal residual sizes computed for each replication is also shown in the plots. These ratios are surprisingly stable.

Compared to the least-squares residuals, the  $L^1$  residuals do fairly well. But, when challenged by a standard robust estimator, they do worse. If

one uses the simulations shown in Figure 2, but replaces the least-squares residuals by robust residuals based on Tukey's biweight with  $6 \times \text{MAD}$ , both produce about an equally large maximal size. However, the ratio of largest to second largest is usually more important for the robust fit, which, therefore, results usually in a clearer picture. This is due to the higher degree of smoothness of such estimators when compared to the  $L^1$  fitter. This also results in an improved relative efficiency.

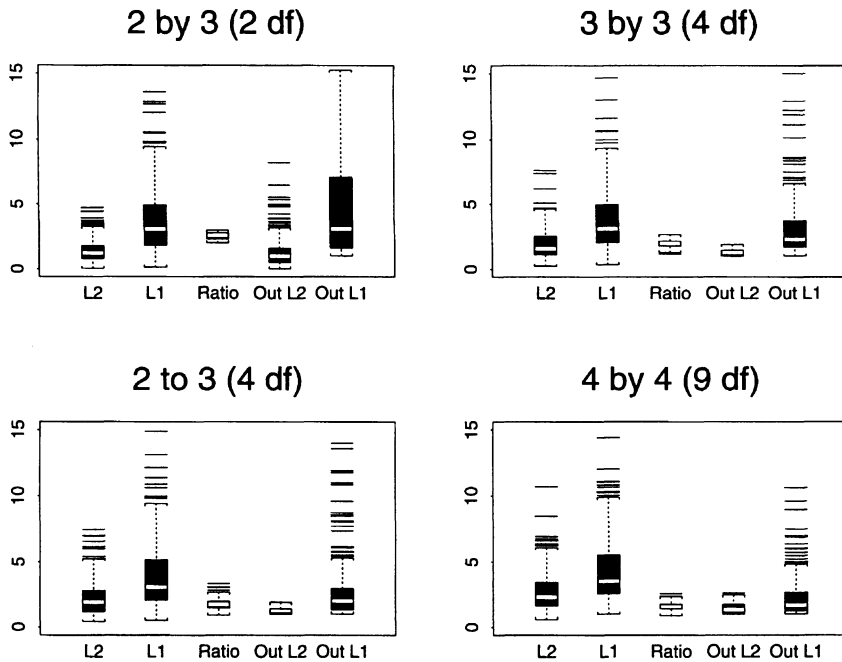


Figure 2: The figure shows the behavior of the maximal residual size in four different design for an error distribution which is Gaussian in all runs except one. In the exceptional one the variance of the Gaussian error is 25 times larger. The average inflation factor observed in 300 replications are 2.61, 1.97, 1.75 and 1.58 and thus remarkably close to the ones observed for Gaussian data. The similarity with the Gaussian case is a bit surprising, since the  $L^1$  method is supposed to be able to identify more clearly the faulty run. In order to check this, each figure contains a boxplot of the ratio of the largest sized residual to the next largest one – the corresponding boxplots are labelled “Out L1” and “Out L2”.

## 4 Non uniqueness of the $L^1$ fit

The ambiguity of the  $L^1$  fit is another problem that the user of this method should be aware of. In the case of the  $2 \times 2$  main effects model, for example, the common  $L^1$  algorithms will result in a residual  $e$  consisting of three exact zeroes and one non-zero residual whose size is exactly four times as large as the size of the  $L^2$  residuals. The run in which the single non zero residual is placed, is completely arbitrary. All intermediate fits preserve the  $L^1$  norm. The same problems occurs in sufficiently balanced replicated  $2 \times 2$  designs such as the one presented in Sheather and McKean (1992, Example 2, p. 153). In their example, the  $L^1$  solution is not unique. In fact, there is a 1-dimensional family of fits, which contains in the center a solution close to the  $L^2$  fit.

In the  $2 \times 3$  design, the  $L^1$  fit results in general in 2 non zero residuals, which are placed in two different columns. The placement of the these two residuals is arbitrary to the extent that we can make them switch rows. All intermediate solutions preserve the  $L^1$  norm. In the  $3 \times 3$  design, things become more complicated. In general, there are four non zero residuals. The values and placements of the non zero cells are usually not unambiguously determined. If the pattern of the non zero cells – indicated by \* – is of the form

|   |   |   |
|---|---|---|
| * | * | 0 |
| * | * | 0 |
| 0 | 0 | 0 |

,

then the fit is unique. If we have, however, a pattern of the form

|   |   |   |
|---|---|---|
| * | 0 | 0 |
| 0 | * | * |
| 0 | * | 0 |

,

there is in general a 2-dimensional set of  $L^1$  solutions.

If we want to recommend the use of  $L^1$  fitting, it seems important to me to produce an algorithm which enumerates all extremal points of the polygone of  $L^1$  fits instead of simply picking one, somewhat at random. This will allow the user to judge, inhowfar the criterion is really identifying outlying points or whether it simply produces large residuals by artificially zeroing others.

## 5 Estimating error variation

The undesirable features of the  $L^1$  residuals that we have discussed above will, of course, have an effect on their ability to predict the error variability.

The  $p$  exact zeroes among the residuals are a property of the design, the fitter and the algorithm. They contain no information about the error distribution. It is, therefore, quite natural to compute the variance of the non zero  $L^1$  residuals as an indication of the variance  $\sigma^2$  of the errors. Consequently, consider an  $L^1$  solution with at least  $p$  zero residuals and at most  $n - p$  non zero residuals. It is evident that the sum of squares of the non zero residuals over-estimates the error variation and the question is by how much. To answer this question, suppose we tried to reconstruct the  $L^2$  residuals  $r$  on the basis of the  $L^1$  residuals  $e$ . If  $p = n - 1$ , then  $e$  would contain a single non-zero residual, which we would evenly distribute over the  $n$   $|r_i|$ . These reconstructed  $L^2$  residuals would then have a sum of squares and, since there is a single degree of freedom, a mean square equal to  $\sum_{i=1}^n e_i^2/n$ . Now, suppose we have several non zero  $L^1$  residuals, which are all of equal size  $|e_i| = R$ . Evenly distributing them over all  $n$  observations, leads to a size of  $|r_i| = R(n-p)/n$ . The mean square of these reconstructed  $r_i$  is equal to  $(n-p)/n^2 \sum_{i=1}^n R^2 = \sum_{i=1}^n e_i^2/n$ , i.e., the same formula as before. One proposal for estimating the error standard deviation from an  $L^1$  fit would, therefore, be the following:

$$u = \sqrt{\sum_{i=1}^n e_i^2/n}.$$

Because of the non-uniqueness,  $\sum_{i=1}^n |e_i|$ , which is uniquely determined, is in some sense a more appropriate basis for estimating  $\sigma$ , the standard deviation of the error. In this case, the over-size of the  $L^1$  residuals does not play any role either. If we re-size them and in some way reconstruct pseudo- $L^2$  residuals, their sum of absolute values would remain the same. How would one estimate  $\sigma$  based on a set of  $L^2$  residuals  $r$ ? Since the marginal distribution of each  $r_i$  has expectation zero and variance  $\sigma^2(1 - h_{ii}) = \sigma^2(n-p)/n$ , we have – for Gaussian errors –  $E(|r_i|) = \sqrt{2/\pi} \sqrt{(n-p)/n} \sigma$ . The statistic

$$\frac{1}{n} \sum_{i=1}^n |r_i| \frac{\pi}{2} \sqrt{n/(n-p)} = \frac{\pi/2}{\sqrt{n(n-p)}} \sum_{i=1}^n |r_i|$$

is, therefore, an unbiased estimate of  $\sigma$ . In replacing  $\sum_{i=1}^n |r_i|$  by  $\sum_{i=1}^n |e_i|$ , we obtain an estimate that underestimates  $\sigma$ , but should still be a useful indicator:

$$v = \frac{\pi/2}{\sqrt{n(n-p)}} \sum_{i=1}^n |e_i|.$$

Figure 3 shows with various plots, how the two estimates behave for Gaussian errors.

A similar behavior is found for other error distributions with a finite second moment.

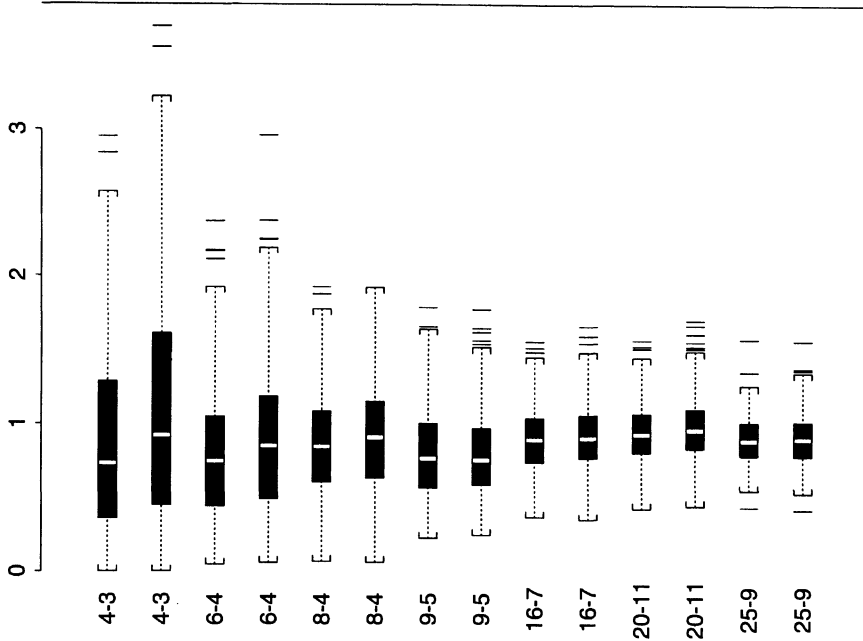


Figure 3: The boxplots show the behavior of the two estimates  $u$  and  $v$  for various balanced designs and Gaussian errors. The labels indicate for each boxplot the number of observations  $n$  and the dimension of the parameter space  $p$ . Each design is represented with two boxplots, the first one for  $u$ , the second one for  $v$ . The true value of  $\sigma$  is equal to 1. The average values over 300 replications are: 0.811, 0.778, 0.843, 0.819, 0.877, 0.968, 0.885 for  $u$  and 1.02, 0.853, 0.888, 0.803, 0.897, 0.984, 0.887 for  $v$ . Both estimates tend to underestimate  $\sigma$ .

## 6 Conclusions

The  $L^1$  method has several drawbacks and in particular leads to some undesirable features built into the residuals. In my opinion it is not a suitable method for fitting of ANOVA data the following reasons:

- (1) The computation of the  $L^1$  fit is not as easy as the computation of the

$L^2$  fit. It is comparable in difficulty to robust fits.

- (2) The  $L^1$  residuals have some idiosyncrasies that should be known to the user of this method. Ignorance will lead to wrong interpretations. They cannot in a straightforward manner replace  $L^2$  residuals.
- (3) The non-uniqueness of the  $L^1$  fit in ANOVA problems is the rule rather than the exception. We lack easily available algorithms which exhibit the whole solution set.
- (4) The resistance of the  $L^1$  fit to outliers is not good enough. One can do better with competing methods that are computationally about equivalent.

## References

- [1] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*. New York: Wiley.
- [2] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.
- [3] Ellis, S.P. and Morgenthaler, S. (1992). Leverage and breakdown in  $L_1$  regression. *J. Am. Statist. Assoc.* **87**, 143–148.
- [4] Müller, Ch. (1995). Outlier robust inference for planned experiments. Habil. Schrift, Fachbereich Mathematik, Freie Universität Berlin.
- [5] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression & Outlier Detection*. New York: Wiley.
- [6] Sheather, S.J. and McKean J.W. (1992). The interpretation of residuals based on  $L_1$  estimation. In  *$L_1$ -Statistical Analysis and Related Methods*, Ed. Y. Dodge. Amsterdam: North-Holland.

# **L<sub>1</sub>-tests in linear models: Tests with maximum relative power**

**Christine Müller**

*Free University of Berlin, Germany*

*Abstract:* In a linear model  $Y_{nN} = x_{nN}^T \beta + Z_{nN}$ ,  $n = 1, \dots, N$ , the problem of testing the hypothesis  $H_0 : L\beta = l$  versus  $H_1 : L\beta \neq l$  is considered. As tests Wald-type tests based on asymptotically linear estimators are used. For such tests the asymptotic efficiency at the ideal model and the asymptotic bias caused by outliers or other deviations from the ideal model depend only on the influence function of the underlying estimator. As for estimation most efficient robust tests can be found by maximizing the efficiency under the side condition that the bias is bounded by some bias bound  $b$ . But this has the disadvantage that the solutions depend on the bias bound  $b$ . To determine  $b$  one can regard measures which are composed by the efficiency and the bias. For estimation such measure is the mean squared error while for testing the power relative to the bias is used. It is shown that the L<sub>1</sub>-tests, i.e. Wald-type tests based on the L<sub>1</sub>-estimator, maximize this relative power. This result is in opposition to that for estimation where the L<sub>1</sub>-estimators do not maximize the mean squared error.

*Key words:* Linear model, L<sub>1</sub>-test, bias of the level, relative power.

AMS subject classification: Primary 62F35 ; secondary 62J05, 62J10, 62K05.

## **1 Introduction**

A general linear model

$$Y_N = X_N \beta + Z_N$$

is considered, where  $Y_N = (Y_{1N}, \dots, Y_{NN})^T$  is the vector of observations,  $\beta \in \mathbb{R}^r$  an unknown parameter vector,  $X_N = (x_{1N}, \dots, x_{NN})^T \in \mathbb{R}^{N \times r}$  the known design matrix with regressors  $x_{1N}, \dots, x_{NN} \in \mathbb{R}^r$  and  $Z_N =$



$(Z_{1N}, \dots, Z_{NN})^T$  the vector of errors. A realization of the random vector  $Y_N$  is denoted by  $y_N = (y_{1N}, \dots, y_{NN})^T$ . In this linear model a hypothesis of the general form

$$H_0 : L\beta = l$$

shall be tested versus the alternative

$$H_1 : L\beta \neq l,$$

where  $l \in \mathbb{R}^s$  and  $\varphi(\beta) = L\beta$  is a linear aspect of the unknown parameter vector  $\beta$  with given matrix  $L \in \mathbb{R}^{s \times r}$  of rank  $s$ .

A large class for testing the hypothesis  $H_0 : L\beta = l$  is the class of Wald-type tests based on asymptotically linear estimators, briefly called ALE-tests (see Müller, 1992a,b, 1995a,b; Rieder, 1994, p.153). To define these tests we assume that the ideal distribution of the standardized errors  $Z_{nN}/\sigma$  is  $P$  and that the design  $x_{1N}, \dots, x_{NN}$  is converging to an asymptotic design measure  $\delta$  in the following sense:

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum e_{x_{nN}}(\{x\}) = \delta(\{x\})$$

for all  $x \in \text{supp}(\delta)$ , where  $\text{supp}(\delta)$  is the support of  $\delta$  and  $e_x$  is the Dirac measure on  $x \in \mathbb{R}^r$ . Then the ALE-tests have a test statistic of the form

$$\tau_N(y_N, X_N) = N (\hat{\varphi}_N(y_N, X_N) - l)^T C_N(y_N, X_N)^{-1} (\hat{\varphi}_N(y_N, X_N) - l),$$

where  $\hat{\varphi}_N$  is an asymptotically linear estimator for  $\varphi(\beta) = L\beta$  with influence function  $\psi$  and  $C(y_N, X_N)$  is a consistent estimator for the asymptotic covariance matrix of  $\hat{\varphi}_N$ , i.e. of  $\sigma^2 C(\psi, \delta)$  with

$$C(\psi, \delta) := \int \psi(z, x) \psi(z, x)^T P(dz) \delta(dx).$$

Thereby an estimator  $\hat{\varphi}_N$  for  $\varphi(\beta) = L\beta$  is called asymptotically linear with influence function  $\psi : \mathbb{R} \times \mathbb{R}^r \rightarrow \mathbb{R}^s$  if  $\int |\psi(z, x)|^2 P(dz) \delta(dx) < \infty$ ,  $\int \psi(z, x) P(dz) = 0$  for all  $x \in \text{supp}(\delta)$ ,  $\int \psi(z, x) x^T z P(dz) \delta(dx) = L$  and

$$\lim_{N \rightarrow \infty} P_{\beta_N}^N \left( -\frac{\sigma}{N} \sum_{n=1}^N \psi \left( \frac{y_{nN} - x_{nN}^T \beta_N}{\sigma}, x_{nN} \right) \middle| > \epsilon \right) = 0$$

for all  $\epsilon > 0$ ,  $\sigma \in \mathbb{R}^+$  and  $\beta_N = \beta + N^{-1/2} \bar{\beta}$  with  $\beta, \bar{\beta} \in \mathbb{R}^r$  and  $L\beta = l$ . The set of all influence functions is denoted by  $\Psi$ . Many wellknown estimators as M-estimators and R-estimators are asymptotically linear.

In particular the classical F-test is an ALE-tests, where  $\hat{\varphi}_N$  is the Gauss-Markov estimator and  $\psi(z, x) = L I(\delta)^{-1} x z$ . Thereby  $M^{-}$  denotes the generalized inverse of the matrix  $M$ , i.e.  $M M^{-} M = M$ , and  $I(\delta)$  the information matrix, i.e.

$$I(\delta) = \int x x^T \delta(dx).$$

If  $\psi(z, x) = \psi^1(z, x) := L I(\delta)^{-1} x \operatorname{sgn}(z) \sqrt{\frac{\pi}{2}}$  then it is the influence function of the L<sub>1</sub>-estimator and the corresponding ALE-test is called L<sub>1</sub>-test.

If the standardized errors  $Z_{1N}/\sigma, \dots, Z_{NN}/\sigma$  are independent and identically distributed according to the ideal distribution  $P$ , then under  $H_0$  the ALE-test statistic  $\tau_N$  has asymptotically a central chi-squared distribution with  $s$  degrees of freedom. Hence, the critical value of an asymptotic level  $\alpha$  ALE-test can be determined as the  $(1-\alpha)$  quantile of the chi-squared distribution. Under contiguous alternatives of the form  $\beta_N = \beta + N^{-1/2}\bar{\beta} \in \mathbb{R}^r$  with  $L\beta_N = l + N^{-1/2}\gamma$  the ALE-test statistic has asymptotically a chi-squared distribution with noncentrality parameter  $\gamma^T [\sigma^2 C(\psi, \delta)]^{-1} \gamma$  so that the power of the test is an increasing function of  $\gamma^T C(\psi, \delta)^{-1} \gamma$ .

If the errors  $Z_{1N}/\sigma, \dots, Z_{NN}/\sigma$  have instead of the ideal distribution  $P$  distributions which are contaminated by outliers and other deviations, then under  $H_0$  the asymptotic error probability can exceed the level  $\alpha$ . The maximum bias of the level, which is possible under contaminated distributions, is an increasing function of

$$\|\psi^T C(\psi, \delta)^{-1} \psi\|_\delta := \max_{(z, x) \in \mathbb{R} \times \operatorname{supp}(\delta)} \psi(z, x)^T C(\psi, \delta)^{-1} \psi(z, x) \quad (1)$$

See Müller (1992a,b, 1995a,b), Rieder (1994), Heritier and Ronchetti (1994).

Let the ideal distribution  $P$  be the standard normal distribution. In Müller (1995a,b) the question was considered which ALE-test has maximum power under the side condition that the maximum bias is bounded by some bias bound. Expressed by influence functions this means: Which  $\psi \in \Psi$  maximizes  $\gamma^T C(\psi, \delta)^{-1} \gamma$  for all  $\gamma \in \mathbb{R}^s$  under the side condition

$$\|\psi^T C(\psi, \delta)^{-1} \psi\|_\delta \leq b, \quad (2)$$

where  $b$  is some given bias bound. But this means that the matrix  $C(\psi, \delta)^{-1}$  should be maximized in the positive definite sense under the side condition (2). In general this optimization problem has no solution (see Krasker and Welsch, 1982). To find solutions one can regard instead of the whole matrix  $C(\psi, \delta)^{-1}$  functionals of the matrix. An appropriate functional for testing is the determinant of the matrix. In Müller (1995a,b) it was shown that for maximizing  $\det(C(\psi, \delta)^{-1})$  under the side condition (2) solutions can be

derived if the design  $\delta$  is D-optimal, i.e.  $\delta \in \arg \min \{\det(L I(\tilde{\delta})^{-1} L^T); \tilde{\delta} \in \Delta\}$ . But the solutions depend on the bias bound  $b$  so that the question exists how to choose  $b$ . Here in Section 2 we propose a criterion for choosing  $b$  based on a relative power value. We show that the best ALE-tests with respect to this criterion is the  $L_1$ -test. In Section 3 we compare this result with a corresponding result for estimation, and in Section 4 we give an example.

## 2 Tests with maximum relative power

Müller (1995a,b) showed that the influence function  $\psi_b$  given by

$$\psi_b(z, x) := \begin{cases} L I(\delta)^{-1} x \operatorname{sgn}(z) \sqrt{\frac{\pi}{2}}, & \text{for } b = s, \\ L I(\delta)^{-1} x \operatorname{sgn}(z) \frac{\min\{|z|, \sqrt{b} y_b\}}{2\Phi(\sqrt{b} y_b) - 1}, & \text{for } b > s, \end{cases}$$

with

$$y_b^2 = \frac{1}{s} g(\sqrt{b} y_b) > 0$$

is a solution of the problem of maximizing the power criterion  $\det(C(\psi, \delta)^{-1})$  under the bias side condition (2) if  $\delta$  is a D-optimal design. Thereby  $\Phi$  denotes the distribution function of the standard normal distribution and  $g$  is given by

$$g(y) := \int \min\{|z|, y\}^2 P(dz).$$

Note that  $\psi_b$  with  $b = s$  is the influence function  $\psi^1$  of the  $L_1$ -test and that  $s = b_{\min} := \min\{\|\psi^T C(\psi, \delta)^{-1} \psi\|_\delta; \psi \in \Psi\}$ . Recall that  $s$  is the rank of the matrix  $L \in \mathbb{R}^{s \times r}$ .

For every solution  $\psi_b$  we have that the quantity (1) providing the maximum bias satisfies

$$\|\psi_b^T C(\psi_b, \delta)^{-1} \psi_b\|_\delta = b$$

and that the power criterion satisfies

$$\det(C(\psi_b, \delta)^{-1}) = \frac{\left(u\left(\frac{b}{s}\right)\right)^s}{\det(L I(\delta)^{-1} L^T)}. \quad (3)$$

Thereby  $u : [1, \infty) \rightarrow (0, \infty)$  is defined by

$$u(a) = \begin{cases} \frac{(2\Phi(\tilde{w}(a)) - 1)^2}{g(\tilde{w}(a))} & \text{for } a > 1, \\ \frac{2}{\pi} & \text{for } a = 1, \end{cases}$$

where  $\tilde{w}(a) > 0$  is implicitly given by  $ag(\tilde{w}(a)) - \tilde{w}(a)^2 = 0$ . In particular we have  $\tilde{w}(\frac{b}{s}) = \sqrt{b}y_b$ . With the implicit function theorem it can be shown that  $u$  is an increasing function (see Müller, 1995a). This means that the power of the ALE test based on  $\psi_b$  increases when the maximum bias value  $b$  increases, and vice versa.

For an appropriate choice of the bias bound  $b$  we can set the power value  $\det(C(\psi_b, \delta)^{-1})$  in relation to the maximum bias given by (1). There are in principle two possibilities:  $b$  should be chosen so that the difference between the power value and the bias value is maximized, or  $b$  should be chosen so that the ratio of the the power value and the bias value is maximal. Maximizing the difference

$$\det(C(\psi_b, \delta)^{-1}) - \|\psi_b^T C(\psi_b, \delta)^{-1} \psi_b\|_\delta = \frac{\left(u\left(\frac{b}{s}\right)\right)^s}{\det(LI(\delta) - L^T)} - b \quad (4)$$

has the disadvantage that the solution would depend on the formulation of the hypotheses. Namely, if we use instead of the hypothese  $H_0 : L\beta = l$  the equivalent hypotheses  $H_0 : \lambda L\beta = \lambda l$  with  $\lambda \neq 1$ , then we have to use  $\tilde{\psi}_b := \lambda \psi_b$  instead of  $\psi_b$ . While the bias value (1) is invariant with respect to  $\lambda$ , this is not the case for the power value so that a solution of maximizing a difference like (4) would depend on  $\lambda$ . This problem does not appear if we use the ratio of the power value and the bias value. Hence,  $b$  should be chosen so that the relative power value

$$\frac{\det(C(\psi_b, \delta)^{-1})^{1/s}}{\|\psi_b^T C(\psi_b, \delta)^{-1} \psi_b\|_\delta} = \frac{1}{b \cdot \det(C(\psi_b, \delta))^{1/s}} \quad (5)$$

is maximized. Thereby we take the  $s$ th root of the determinant of the covariance matrix to ensure that an improvement of the covariance matrix by a factor  $c$  provides also an improvement of the relative power value by the factor  $c$ . Note also that the  $s$ th root of the determinant is often used as a measure for the entropy and that it is the geometric mean of the diagonal elements if the covariance matrix is a diagonal matrix.

The following theorem shows that the L1-test, i.e. the ALE-test based on  $\psi_b$  with  $b = s$ , has the maximum relative power.

**Theorem 1**  *$b = s$  maximizes the relative power value (5) with respect to  $b$ , i.e. the  $L_1$ -test has maximum relative power.*

**Proof:** Using (3) maximizing of (5) is equivalent to minimizing

$$\begin{aligned} b \cdot \det(C(\psi_b, \delta))^{1/s} &= b \cdot \frac{1}{u\left(\frac{b}{s}\right)} \det(LI(\delta) - L^T)^{1/s} \\ &= t\left(\frac{b}{s}\right) \cdot s \cdot \det(LI(\delta) - L^T)^{1/s} \end{aligned}$$

where

$$t(a) := a \frac{1}{u(a)} \text{ for } a \geq 1.$$

By calculating the first and second derivative of  $t : [1, \infty) \rightarrow (0, \infty)$  it can be shown that  $t$  is a convex function with  $\lim_{a \downarrow 1} t'(a) = 0$ , where  $t'$  denotes the first derivative of  $t$ . This means that  $t\left(\frac{b}{s}\right)$  is minimized by  $b = s$ . To calculate the first and the second derivative of  $t$  it is helpful to calculate the derivatives of  $u$ . At first note that the implicit function theorem provides

$$\tilde{w}'(a) = \frac{\tilde{w}(a) g(\tilde{w}(a))}{2a [2\Phi(\tilde{w}(a)) - 1 - 2\tilde{w}(a) \Phi'(\tilde{w}(a))]} > 0.$$

Then, with  $h(y) := y\Phi(-y) - \Phi'(y)$ , it can be shown that

$$\begin{aligned} u'(a) &= \frac{-(2\Phi(\tilde{w}(a)) - 1) 2h(\tilde{w}(a))}{\tilde{w}(a)} > 0, \\ u''(a) &< 0 \text{ and} \\ u''(a)a + 2u'(a) &\leq 0 \end{aligned}$$

is satisfied for all  $a > 1$  (see Müller, 1995a, Lemma 12.7). The rule of L'Hospital provides  $\lim_{a \downarrow 1} \tilde{w}(a) = 0$  and  $\lim_{a \downarrow 1} u(a) = \frac{2}{\pi}$  so that

$$t'(a) = \frac{1}{u(a)} \left( 1 + \frac{2h(\tilde{w}(a))\tilde{w}(a)}{2\Phi(\tilde{w}(a)) - 1} \right)$$

is converging to 0 for  $a \downarrow 1$  (see Müller 1995a, Lemma 12.8).  $\square$

### 3 Comparison with estimation problems

An asymptotically linear estimator  $\hat{\varphi}_N$  for estimating  $\varphi(\beta) = L\beta$  with influence function  $\psi$  is under contaminated distributions asymptotically normally distributed. The asymptotic covariance matrix is  $C(\psi, \delta)$  and the maximum asymptotic bias is given by

$$\|\psi\|_\delta := \max_{(z,x) \in \mathcal{R} \times \text{supp}(\delta)} |\psi(z, x)| \quad (6)$$

(see Bickel, 1981, 1984; Rieder, 1985, 1987, 1994). Solutions  $\psi_b^e$  which minimize the trace of the covariance matrix  $C(\psi, \delta)$  under the bias side condition  $\|\psi\|_\delta \leq b$  can be characterized explicitly if the design  $\delta$  is based on linearly independent regressors or if the design  $\delta$  is A-optimal, i.e.  $\delta \in \arg \min \{\text{tr}(L I(\tilde{\delta})^{-1} L^T); \tilde{\delta} \in \Delta\}$  (see Kurotschka and Müller, 1992; Müller, 1994a). As for testing the optimal influence function  $\psi_b^e$  depends heavily

on the bias bound  $b$ . Moreover, only the solution  $\psi_b^e$  with  $b = b_{\min}^e := \min\{\|\psi\|_\delta; \psi \in \Psi\}$  corresponds to the L<sub>1</sub>-estimator, i.e. satisfies  $\psi_b^e = \psi^1$ .

Similarly to testing an optimal bias bound can be found by combining the efficiency criterion  $\text{tr}(C(\psi_b^e, \delta))$  with the bias value  $\|\psi_b^e\|_\delta$ . A natural combined criterion is the asymptotic mean squared error

$$M_\delta(b) := \|\psi_b^e\|_\delta^2 + \text{tr}(C(\psi_b^e, \delta)) = b^2 + \text{tr}(C(\psi_b^e, \delta)). \quad (7)$$

It can be shown that as for testing  $M_\delta$  is convex. But in opposition to testing  $M_\delta$  attains its minimum for a value  $b > b_{\min}^e$  so that the L<sub>1</sub>-estimator does not minimize the mean squared error. See Müller (1994b,c).

Hence, we have the following situation: The approaches for estimation and testing look very similar and leads to similar constrained optimization problems of maximizing the efficiency under a bias bound  $b$ . Nevertheless the problem of finding an optimal bias bound by using a natural criterion which combines efficiency and bias leads to qualitative different results. For testing the best bias bound is  $b = b_{\min} = s$  so that the L<sub>1</sub>-test is optimal while for estimation the best bias bound is  $b > b_{\min}^e$  so that the L<sub>1</sub>-estimator is not optimal. Moreover, for estimation the optimal bias bound depends strongly on the model, the design and the aspect  $L\beta$  and it can be calculated only per computer. For testing the optimal bias bound is simply  $s$ , the rank of  $L$ .

## 4 Example

Consider a one-way lay-out model with four levels. i.e. we have four samples with unknown means  $\beta_1, \beta_2, \beta_3, \beta_4$  so that the observations are given by

$$Y_{nN} = \beta_i + Z_{nN},$$

if the observation  $Y_{nN}$  belongs to the sample  $i$ ,  $i = 1, 2, 3, 4$ . This model can be expressed as a linear model with  $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^T \in \mathbb{R}^4$  and

$$x_{nN} = x(i) := (1_1(i), 1_2(i), 1_3(i), 1_4(i))^T.$$

Often one sample, say sample 1, is a controll group. Then an interesting aspect of  $\beta$  is the linear aspect  $\varphi(\beta) = (\beta_2 - \beta_1, \beta_3 - \beta_1, \beta_4 - \beta_1)^T \in \mathbb{R}^3$ . Then testing  $H_0 : \varphi(\beta) = 0$  against  $H_1 : \varphi(\beta) \neq 0$  is equivalent with testing  $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ . A D-optimal design for  $\varphi(\beta)$  is  $\delta = \frac{1}{4}(e_{x(1)} + e_{x(2)} + e_{x(3)} + e_{x(4)})$  which means that the four samples are of equal size. At this design the influence function of the L<sub>1</sub>-test for  $H_0$  and the L<sub>1</sub>-estimator for  $\varphi(\beta)$  has the form

$$\psi^1(z, x(i)) = \begin{cases} (-1, -1, -1)^T \text{sgn}(z) 4 \sqrt{\frac{\pi}{2}}, & \text{for } i = 1, \\ (1_2(i), 1_3(i), 1_4(i))^T \text{sgn}(z) 4 \sqrt{\frac{\pi}{2}}, & \text{for } i \neq 1. \end{cases}$$

According to Theorem 1 this influence function provide the maximum relative power within all ALE-tests for  $H_0$ . But this influence function does not provide the minimum mean squared error within all asymptotically linear estimators for  $\varphi(\beta)$ . The influence function providing the minimum mean squared error has the form

$$\psi_b^e(z, x(i)) = \begin{cases} (-1, -1, -1)^T \operatorname{sgn}(z) \frac{\min\{|z|, b w_b\}}{w_b \sqrt{3}}, & \text{for } i = 1, \\ (1_2(i), 1_3(i), 1_4(i))^T \operatorname{sgn}(z) \frac{\min\{|z|, b v_b\}}{v_b}, & \text{for } i \neq 1, \end{cases}$$

where  $b \approx 8.7213$ ,  $w_b \approx 0.0186$  and  $v_b \approx 0.2411$  (see Müller, 1994c).

## References

- [1] Bickel, P.J. (1981). Quelques aspects de la statistique robuste. In *Ecole d'Eté de Probabilités de St. Flour*, Springer Lecture Notes in Math, Vol. 876, pp. 1-72.
- [2] Bickel, P.J. (1984). Robust regression based on infinitesimal neighbourhoods. *Ann. Statist.* **12**, 1349-1368.
- [3] Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *J. Am. Statist. Assoc.* **89**, 897-904.
- [4] Krasker, W.S. and Welsch, R.E. (1982). Efficient bounded-influence regression estimation. *J. Am. Statist. Assoc.* **77**, 595-604.
- [5] Kurotschka, V. and Müller, Ch.H. (1992). Optimum robust estimation of linear aspects in conditionally contaminated linear models. *Ann. Statist.* **20**, 331-350.
- [6] Müller, Ch.H. (1992a).  $L_1$ -estimation and testing in conditionally contaminated linear models. In  *$L_1$ -Statistical Analysis and Related Methods*, Ed. Y. Dodge, pp. 69-76. Amsterdam: North-Holland.
- [7] Müller, Ch.H. (1992b). Robust testing in conditionally contaminated linear models. Preprint No. A-92-12, Freie Universität Berlin, Fachbereich Mathematik und Informatik.
- [8] Müller, Ch.H. (1994a). Optimal designs for robust estimation in conditionally contaminated linear models. *J. Statist. Plann. Infer.* **38**, 125-140.
- [9] Müller, Ch.H. (1994b). Optimal bias bounds for robust estimation in linear models. In *Proc. of the International Conference on Linear Statistical Inference LINSTAT'93*, Eds. T. Caliński and R. Kala, pp. 97-102. Dordrecht: Kluwer Academic Publishers.
- [10] Müller, Ch.H. (1994c). On the calculation of MSE minimizing robust estimators. In *COMPSTAT'94, Proceedings in Computational Statis-*

*tics*, Eds. R. Dutter and W. Grossmann, pp. 257-262. Heidelberg: Physica-Verlag.

- [11] Müller, Ch.H. (1995a). Outlier robust inference for planned experiments. Habilitation thesis, Freie Universität Berlin. To appear as Springer Lecture Notes in Statistics.
- [12] Müller, Ch.H. (1995b). Optimum robust testing in linear models. Preprint No. A-28-95, Freie Universität Berlin, Fachbereich Mathematik und Informatik. Submitted to *Ann. Statist.*
- [13] Rieder, H. (1985). Robust estimation of functionals. Technical Report, Universität Bayreuth.
- [14] Rieder, H. (1987). Robust regression estimators and their least favorable contamination curves. *Stat. Decis.* **5**, 307-336.
- [15] Rieder, H. (1994). *Robust Asymptotic Statistics*. New York: Springer.





# On $L_1$ -norm estimators in nonlinear regression and in nonlinear error-in-variables models

Silvelyn Zwanzig

*University of Hamburg, Germany*

*Abstract:* Strong consistency results for the  $L_1$ -norm estimator of the parameter of interest are shown in nonlinear models, namely in the nonlinear regression model, in the nonlinear error-in-variables model and in a nonlinear semiparametric model.

*Key words:* Nonlinear semiparametric regression, nonlinear functional relation, minimum contrast estimates,  $L_1$ -norm, consistency.

AMS subject classification: 62F12.

## 1 Introduction

In regression models the  $L_1$ -norm estimators receive their justification from robustness theory. For the error distribution only assumptions on the behavior around the median, which should be zero, are required for consistency results.

For nonlinear regression the consistence of the  $L_1$ -norm estimators is shown by Oberhofer (1982). Richardson and Bhattacharyya (1987) extended this result to general noncompact parameter sets by using a sieves method. The general approach of M-estimators in Liese and Vajda (1994) for nonlinear regression models includes also the  $L_1$ -norm estimator. They obtained a similar result as Richardson and Bhattacharyya (1987) with a different method and conditions that are statistically more transparent.

The concept of minimum contrast estimators and the method of sieves are studied in the nonparametric theory by van de Geer (1990), van de Geer (1995), Birge and Massart (1991) and Birge and Massart (1994). They

apply their general results for consistency rates of minimum contrast estimators to the  $L_1$ -norm estimator of the nonparametric regression function. (Compare also the Section 3.4.4. in the recent book of van der Vaart and Wellner, 1996).

The aim of this paper is to give consistency results for  $L_1$ -norm estimator in semiparametric regression models, where the parameter of interest is of fixed dimension and the nuisance parameter is either given as an unknown function from a nonparametric function space or has a dimension which increases with the sample size. In this paper we will study simultaneously three models, namely nonlinear regression, nonlinear functional relation, nonlinear semiparametric regression, all to be introduced in Section 2. This allows us to demonstrate the line of proving the  $L_1$ -norm consistency results and to emphasize the underlying problems. For illustration reasons the  $L_1$ -approach is also embedded in the minimum contrast context, see Section 3. The proof consists mainly of two steps given in Section 4: first the approximation of the empirical  $L_1$ -norm by its expected value, which will be done in Subsection 4.1, and second the identification of the parameter by using the expected values of the difference of empirical  $L_1$ -norms in Subsection 4.2. The identification problem is specific for the individual models. This is not characteristic for the  $L_1$ -norm approach, since the same problem also occurs in the  $L_2$ -theory. The technique of approximation is more or less standard and is based on results of the increments of sub-Gaussian processes. In Section 5 the consistency results for each model are separately given. For the nonlinear regression model this is a known strong consistency result in form of an exponential probability inequality. The consistency result for the nonlinear functional relation models is new. Under this kind of entropy condition on the nuisance parameter space only results are known for the least squares estimator, see Zwanzig (1990). The nonlinear semiparametric model is of a special structure, because of the constraints imposed by the identification problem. Linton (1995) studied this model with a linear parametric part. The result given here for this model seems to be new as well.

## 2 The general setting

Here we introduce a general semiparametric regression model to be specified in the following. Suppose we have independent and in general not identically distributed two dimensional real valued observations  $(y_1, x_1), \dots, (y_n, x_n)$ , generated by

$$y_i = g(\xi_i, \beta) + \varepsilon_{1i}, \quad (1)$$

$$x_i = h(\xi_i, \beta) + \varepsilon_{2i}, \quad (2)$$

with  $i = 1, \dots, n$ . The probability of each observation  $(y_i, x_i)$  is  $P_{\xi_i \beta}$ . The common distribution of the whole sample is denoted by  $P_{\xi \beta} = \prod_{i=1}^n P_{\xi_i \beta}$  and dominated by a  $\sigma$ -finite measure  $\mu_n$ .

The errors  $\varepsilon_{ji}$  are i.i.d. with distribution  $P_\varepsilon$ , expected value zero and positive variances  $\sigma^2$ . The error distribution does not depend of the parameter  $\beta$ . To get the consistency of  $L_1$ -norm estimators we will need the assumption **E** on the error distribution that the median is zero and that the distribution has enough mass in the local neighborhood of it:

**E**  $\exists D_0 \exists \kappa_\varepsilon \forall d \leq D_0$  such that

$$P_\varepsilon(-d \leq \varepsilon \leq 0) \geq \kappa_\varepsilon d \quad , \quad P_\varepsilon(0 \leq \varepsilon \leq d) \geq \kappa_\varepsilon d. \quad (3)$$

$$P_\varepsilon(\varepsilon \geq 0) = \frac{1}{2}$$

The functions  $h(.,.)$  and  $g(.,.)$  are continuous and known. The regression parameter  $\beta \in \mathcal{B} \subset \mathbb{R}^p$  is the parameter of interest. The dimension  $p$  of  $\beta$  does not depend on the sample size  $n$ . The design points or variables  $\{\xi_1, \dots, \xi_n\} \subset \mathbb{R}$  are unknown and fixed. They are the nuisance parameters, whose number grows with the sample size  $n$ . We write the nuisance parameters as components of a column vector of dimension  $n$ :  $\xi = \xi^{(n)} = (\xi_1, \dots, \xi_n)^T \in \mathcal{X}^{(n)} \subseteq \mathbb{R}^n$ . The common unknown parameter is

$$\theta = (\xi, \beta) \in \Theta = \mathcal{X}^{(n)} \times \mathcal{B} \subseteq \mathbb{R}^{n+p}. \quad (4)$$

The model assumptions (1), (2) above include different models for different specification of  $\mathcal{X}^{(n)} \subseteq \mathbb{R}^n$  and of the functions  $h$  and  $g$ .

## 2.1 The nonlinear regression model

Suppose the design is known. That is, we have  $\mathcal{X}^{(n)} = \{\xi^0 = (\xi_1^0, \dots, \xi_n^0)^T\}$ . We consider only the first equation (1) and obtain the nonlinear regression model with  $n$  observations

$$y_i = g(\xi_i^0, \beta) + \varepsilon_{1i}, \quad \text{with} \quad i = 1, \dots, n. \quad (5)$$

Note that in this model no nuisance parameters occur.

## 2.2 The nonlinear error-in-variables model

For  $h(\xi_i, \beta) = \xi_i$  in (2) we obtain the nonlinear error-in-variables model,

$$y_i = g(\xi_i, \beta) + \varepsilon_{1i}, \quad (6)$$

$$x_i = \xi_i + \varepsilon_{2i}, \quad (7)$$

with  $i = 1, \dots, n$ . This is exactly the functional one, because the variables are considered as fixed and unknown and play the role of a nuisance parameter with increasing dimension. For consistency we need additional assumptions on the set of nuisance parameters  $\mathcal{X}^{(n)}$ , because we know that for  $\mathcal{X}^{(n)} = [0, 1]^n$  the least squares estimator is inconsistent, see Kukush and Zwanzig (1996). One interesting additional information may be

$$\mathcal{X}^{(n)} = \left\{ \xi = (\xi_1, \dots, \xi_n)^T : 0 \leq \xi_1 \leq \xi_2 \leq \dots \leq \xi_{n-1} \leq \xi_n \leq 1 \right\}. \quad (8)$$

On the first view this assumption seems to be artificial; it is, however, useful in applications, for instance in biology or chemistry. There the unknown design points  $\xi_i$  often stand for different levels of concentration. The experimenter measures these concentration levels with error  $\varepsilon_{2i}$ . But he does have some influence on the level that the concentration lies at and he can guarantee with high security that the concentration level of the next experiment will be higher.

The assumption (8) can be rewritten

$$\mathcal{X}^{(n)} = \left\{ \xi = (f(z_1), \dots, f(z_i), \dots, f(z_n))^T : f : [0, 1] \rightarrow [0, 1], f \text{ increasing} \right\}. \quad (9)$$

where the  $z_1 \leq \dots \leq z_n$  are fixed design points satisfying the following design condition:

**D**

$$\lim_{n \rightarrow \infty} \max_i |z_i - z_{i-1}| = 0.$$

Another possibility is to consider

$$\mathcal{X}^{(n)} = \left\{ \xi = (f(z_1), \dots, f(z_i), \dots, f(z_n))^T : f \in \mathcal{X}_{m,\alpha}(C, L) \right\} \quad (10)$$

with

$$\mathcal{X}_{m,\alpha}(C, L) = \left\{ f \in C_{m,\alpha}[0, 1] : \begin{array}{l} |f^{(k)}(x)| \leq C, k = 0, 1, \dots, m \\ |f^{(m)}(x_1) - f^{(m)}(x_2)| \leq L |x_1 - x_2|^\alpha \end{array} \right\}. \quad (11)$$

This kind of additional information is also used in the following semi-parametric model.

### 2.3 The nonlinear semiparametric model

Assume that  $\{z_1, \dots, z_n\} \subset [0, 1]$  are known design points with **D**. Under the specification  $\xi_i = f(z_i)$  and  $g(\xi_i, \beta) = m(z_i, \beta) + \xi_i$ , where the functions  $m$  and  $f$  satisfy the following identification condition,

**ID**  $\exists n_0 \forall n > n_0 \exists \{w_i\} > 0, \sum_{i=1}^n w_i = 1 \forall f \in \mathcal{X}_{m,\alpha}(C, L) \forall \beta \in \mathcal{B}^c$  such that

$$\sum_{i=1}^n w_i m(z_i, \beta) f(z_i) = 0, \quad (12)$$

we consider only the first equation (1) and obtain the model

$$y_i = m(z_i, \beta) + f(z_i) + \varepsilon_{1i}. \quad (13)$$

The condition **ID** contains the orthogonality in the sense of the empirical measure generated by the weighted design points  $z_1, \dots, z_n$ . This model (13) describes alternatives in the context of model choice.

## 3 Minimum contrast estimates

The  $L_1$ -norm estimator will be considered as a special minimum contrast estimator. We call a nonrandom positive real function  $C_n : \theta \in \Theta^c \rightarrow \mathbb{R}_+$  a *contrast for  $\theta$  at  $\theta_{(n)}$*  iff it is lower semicontinuous and

$$\theta_{(n)} = \arg \min_{\theta \in \Theta^c} C_n(\theta), \quad (14)$$

where  $\Theta^c$  denotes the compactification of the parameter set  $\mathcal{X}^{(n)} \times \mathcal{B}$  in  $\overline{\mathbb{R}}^{n+p}$ . The contrast may depend on the sample size  $n$ . Examples are the empirical  $L_q$ -contrast

$$C_n(\theta) = \sum_{i=1}^n w_i \left| g(\xi_i^0, \beta) - g(\xi_i^0, \beta^0) \right|^q \quad (15)$$

or the asymptotic  $L_q$ -contrast

$$C_n(\theta) = C(\beta) = \int \left| g(x, \beta) - g(x, \beta^0) \right|^q dG, \quad (16)$$

with the  $\theta_{(n)} = (\beta^0, \xi^0)$  satisfying (1) and (2). The first depends on the unknown design points  $\xi \in \mathcal{F}^{(n)}$ , the other on an asymptotic design  $G$ . Under a unique parameterization of the regression function each distance

measure  $d$  for functions  $g(., \beta) \in \mathcal{M}$  seems to be a useful contrast at  $\beta_{(n)} = \beta^0$ , namely  $C_n(\beta) = d(g(., \beta), g(., \beta^0))$ .

We call a measurable function

$$\tilde{C}_n(., ., .) : \mathbb{R}^n \times \mathbb{R}^n \times \Theta^c \rightarrow \mathbb{R}_+ \quad (17)$$

a *contrast function* and require that it is continuous with respect to  $\theta$ . In the general statistical experiment, which includes random processes and random fields as well, Liese and Vajda (1995) introduced a more general concept and called the function corresponding to (17) a contrast principle. Note in order to simplify the denotation we will suppress the dependence on the sample and let the tilde hints to this:  $\tilde{C}_n(X, Y, \theta) =: \tilde{C}_n(\theta)$ . We then define the corresponding estimator as follows.

A measurable solution  $\tilde{\theta} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \Theta^c$  is called a *minimum contrast estimator* iff

$$\tilde{\theta} \in \arg \min_{\theta \in \Theta^c} \tilde{C}_n(\theta). \quad (18)$$

Under the model assumptions above the existence of minimum contrast estimators are given by the Lemma 2 of Liese and Vajda (1995).

The following lemma gives the connection between the consistency of the minimum contrast estimator and the uniform consistent approximation of the contrast  $C_n(\theta)$  by the contrast function  $\tilde{C}_n(\theta)$ . It is a version of an “argmin” result, like the argmax theorem for i.i.d. experiments in van der Vaart and Wellner (1996). Consider the differences of the contrast and of the contrast function,

$$\Delta C_n(\theta) = C_n(\theta) - C_n(\theta_{(n)}) \quad \text{and} \quad \Delta \tilde{C}_n(\theta) = \tilde{C}_n(\theta) - \tilde{C}_n(\theta_{(n)}). \quad (19)$$

**Lemma 1** *Let  $\rho : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  be strictly increasing, with  $\rho(0) = 0$ . Let  $d(., .)$  be a semimetric on  $\Theta^c$ . For any  $\epsilon > 0$  define the set*

$$\bar{\Theta}_n(\epsilon) = \Theta^c \cap \left\{ \theta : d(\theta, \theta_{(n)}) > \epsilon \right\}. \quad (20)$$

*Let  $C_n$  be a contrast with*

$$\Delta C_n(\theta) = C_n(\theta) - C_n(\theta_{(n)}) \geq \rho(d(\theta, \theta_{(n)})). \quad (21)$$

*Then*

$$\forall \epsilon > 0 \quad P(d(\tilde{\theta}, \theta_{(n)}) > \epsilon) \leq P\left(\sup_{\theta \in \bar{\Theta}_n(\epsilon)} \frac{|\Delta C_n(\theta) - \Delta \tilde{C}_n(\theta)|}{\rho(d(\theta, \theta_{(n)}))} \geq 1\right). \quad (22)$$

**Proof:** From (21) follows that

$$\min_{\theta \in \bar{\Theta}_n(\epsilon)} \frac{\Delta C_n(\theta)}{\rho(d(\theta, \theta_{(n)}))} \geq 1. \quad (23)$$

Under  $\tilde{\theta} \in \bar{\Theta}_n(\epsilon)$  we have  $\tilde{C}_n(\tilde{\theta}) \leq \tilde{C}_n(\theta_{(n)})$  and  $\rho(d(\tilde{\theta}, \theta_{(n)})) > \rho(\epsilon) > 0$ , thus

$$\frac{\Delta \tilde{C}_n(\tilde{\theta})}{\rho(d(\tilde{\theta}, \theta_{(n)}))} \leq 0. \quad (24)$$

Hence under  $\tilde{\theta} \in \bar{\Theta}_n(\epsilon)$  we obtain from (23) and (24) the following chain of inequalities

$$1 \leq \min_{\theta \in \bar{\Theta}_n(\epsilon)} \frac{\Delta C_n(\theta)}{\rho(d(\theta, \theta_{(n)}))} - \frac{\Delta \tilde{C}_n(\tilde{\theta})}{\rho(d(\tilde{\theta}, \theta_{(n)}))} \leq \sup_{\theta \in \bar{\Theta}_n(\epsilon)} \frac{|\Delta C_n(\theta) - \Delta \tilde{C}_n(\theta)|}{\rho(d(\theta, \theta_{(n)}))}, \quad (25)$$

which yields the statement of the lemma.  $\square$

Note that the rate of convergence of the minimum contrast estimator given by this lemma depends mainly on the separation property of the contrast (21) and the semimetric  $d(.,.)$  chosen in (21).

## 4 The $L_1$ -estimator

In the following we focus our attention on the  $L_1$ -contrast function only. This means, let from now on

$$\Delta \tilde{C}_n(\theta) = \sum_{i=1}^n w_{1i} (|\varepsilon_{1i} + \Delta g(\xi_i, \beta)| - |\varepsilon_{1i}|) + \sum_{i=1}^n w_{2i} (|\varepsilon_{2i} + \Delta h(\xi_i, \beta)| - |\varepsilon_{2i}|) \quad (26)$$

with

$$\Delta g(\xi_i, \beta) = g(\xi_i^0, \beta^0) - g(\xi_i, \beta) \text{ and } \Delta h(\xi_i, \beta) = h(\xi_i^0, \beta^0) - h(\xi_i, \beta).$$

We will see that the  $L_1$ -contrast is

$$\Delta C_n(\theta) = E_{\xi^0 \beta^0} \Delta \tilde{C}_n(\theta). \quad (27)$$

It is easily checked that the  $L_1$ -contrast does not coincide with the contrast defined by the empirical  $L_1$ -metric (15) or by the asymptotic  $L_1$ -metric (16), see Lemma 4.

Applying Lemma 1 for the consistency proof of the  $L_1$ -estimator, we have to do two steps: first the uniform approximation of the contrast by the contrast function with an appropriate rate, second the study of the separation condition (21).



### 4.1 The approximation

One of the main advantages of the  $L_1$ -approach is that the difference

$$Z(\theta) = \Delta C_n(\theta) - \Delta \tilde{C}_n(\theta) \quad (28)$$

as a stochastic process  $\{Z(\theta), \theta \in \Theta^c\}$  with index set  $\Theta^c \subseteq \bar{\mathbb{R}}^{n+p}$  forms an sub-Gaussian process without additional assumptions on the tails of the error distribution  $P_\varepsilon$ . The only things we need are Lipschitz conditions and a convenient semimetric  $d$  in  $\Theta^c$ . Henceforth, we will use the following denotation for the sum of weighted squares with known normalized weights

$$w_{ji} > 0, \quad \sum_{i=1}^n w_{ji} = 1, \quad w_{\max} = \max_{ij} w_{ji} \quad (29)$$

$$\sum_{i=1}^n w_{1i} |g(\xi_i, \beta)|^2 = \int |g(x, \beta)|^2 dG_{w_1}(\xi) = |g(\xi, \beta)|_{w_1}^2,$$

where  $G_{w_1}(\xi)$  is the weighted empirical measure generated by the design points  $\xi \in \mathcal{X}^{(n)}$ . We use also the corresponding notation for the scalar product  $(\cdot, \cdot)_{w_1}$ . Note that in the unweighted case  $w_{\max} = n^{-1}$ . The Lipschitz conditions are used with respect to both types of parameters.

**L1**  $\exists n_0 \exists L_1, L_1 < \infty \forall n \geq n_0 \forall \beta \in \mathcal{B}^c \forall \xi, \xi' \in \mathcal{X}^{(n)c}$ , such that

$$|g(\xi, \beta) - g(\xi', \beta)|_{w_1}^2 + |h(\xi, \beta) - h(\xi', \beta)|_{w_2}^2 \leq L_1 \|\xi - \xi'\|_{w_2}^2. \quad (30)$$

**L2**  $\exists n_0 \exists L_2, L_2 < \infty, \forall n \geq n_0 \forall \xi \in \mathcal{X}^{(n)c} \forall \beta, \beta' \in \mathcal{B}^c$ , such that

$$|g(\xi, \beta) - g(\xi, \beta')|_{w_1}^2 + |h(\xi, \beta) - h(\xi, \beta')|_{w_2}^2 \leq L_2 \|\beta - \beta'\|^2. \quad (31)$$

In (31)  $\|\cdot\|$  denotes the Euclidean norm in  $\mathbb{R}^p$ .

**Lemma 2** Suppose **L1**, **L2**. Set

$$d(\theta, \theta')^2 = w_{\max} \left[ \|\xi - \xi'\|_{w_2}^2 + \|\beta - \beta'\|^2 \right]. \quad (32)$$

Then there is a constant  $\alpha = \alpha(L_1, L_2)$  independent of  $n$  and  $\theta$  such that for all  $t \geq 0$  and all  $n \geq n_0$

$$P_\varepsilon(|Z(\theta) - Z(\theta')| \geq t) \leq 2 \exp\left(-\frac{\alpha t^2}{d(\theta, \theta')^2}\right). \quad (33)$$

**Proof:** We have for  $Z(\theta)$  as defined in (28) with (26), (27) that  $Z_n(\theta) = \sum_{i=1}^n X_i(\theta)$  and  $EX_i(\theta) = 0$ . Using the inequality  $||a+b| - |a+c|| \leq |b+c|$  we obtain

$$|X_i(\theta) - X_i(\theta')| \leq 2w_{1i} |g(\xi_i, \beta) - g(\xi'_i, \beta')| + 2w_{2i} |h(\xi_i, \beta) - h(\xi'_i, \beta')|,$$

and using the triangle inequality we get further

$$|X_i(\theta) - X_i(\theta')| \leq d_i(\theta, \theta') \quad (34)$$

where  $d_i(\theta, \theta')$  denotes the following seminorm  $d_i(\theta, \theta') = d_{1i}(\theta, \theta') + d_{2i}(\theta, \theta')$ , with

$$d_{1i}(\theta, \theta') = 2w_{1i} |g(\xi_i, \beta) - g(\xi'_i, \beta)| + 2w_{1i} |g(\xi'_i, \beta) - g(\xi'_i, \beta')|,$$

$$d_{2i}(\theta, \theta') = 2w_{2i} |h(\xi_i, \beta) - h(\xi'_i, \beta)| + 2w_{2i} |h(\xi'_i, \beta) - h(\xi'_i, \beta')|.$$

Because of (34), we can apply the Corollary 3.2 of van de Geer (1990). Thus there is a constant  $\alpha'$  independent of  $n$  and  $\theta$  such that

$$P_{\xi^0 \beta^0} (|Z(\theta) - Z(\theta')| \geq t) \leq 2 \exp \left( -\frac{t^2 \alpha'}{2\bar{d}(\theta, \theta')^2} \right).$$

with  $\bar{d}(\theta, \theta')^2 = \sum_{i=1}^n d_i(\theta, \theta')^2$ . Under the conditions **L1**, **L2** we estimate

$$\sum_{i=1}^n d_i(\theta, \theta')^2 \leq 16w_{\max} (L_1 + L_2) \left( |\xi - \xi'|_{w_2}^2 + \|\beta - \beta'\|^2 \right)$$

and obtain the statement with  $\alpha = \alpha' (32(L_1 + L_2))^{-1}$ .  $\square$

In order to formulate the entropy condition we need a few more definitions. Let us introduce them for a general set  $A$  with a metric  $d$ , because inside the proof we will use the notion of entropy in the context of several different sets. A family of subsets  $U_1, \dots, U_N$  is called an  $\epsilon$ -covering of  $A$  with respect to a metric  $d$ , if the diameter of each  $U_k$  does not exceed  $2\epsilon$  and if the sets cover  $A$ ,  $A \subseteq \cup_i^N U_i$ . The  $\epsilon$ -covering number  $N(\epsilon)$  is the minimal number of  $U_k$ 's in any  $\epsilon$ -covering of  $A$ . The  $\epsilon$ -entropy  $H(\epsilon)$  of  $A$  is given by the logarithm  $H(\epsilon) = \ln N(\epsilon)$ . The entropy depends on the metric  $d$  and on  $A$ . We therefore denote the local  $\epsilon$ -entropy of  $A \cap (a : d(a, a_0) \leq D)$  by  $H_{A,d}(\epsilon, D)$ . We will require a condition on the local entropy of the nuisance parameter set only, that is  $A = \mathcal{X}^{(n)}$  and  $d(\xi, \xi') = |\xi - \xi'|_{w_2}$ . For abbreviation we write  $H_{\mathcal{X}^{(n)}, |\cdot|_{w_2}}(\epsilon, D) = H_\xi(\epsilon, D)$ . The entropy condition we need is:

**Ent** For all  $\delta, \eta_n > 0$ , with  $\delta \eta_n \sqrt{w_{\max}} \geq 1$

$$\lim_{L \rightarrow \infty} \frac{\int_0^1 \sqrt{H_\xi(Lu\delta, L\delta D)} du}{L\delta \eta_n \sqrt{w_{\max}}} = 0. \quad (35)$$

The following lemma is an application of a modified result of van de Geer (1990), which is an adaptation of the chaining method of Pollard (1984), on page 144.

**Lemma 3** Suppose **L1** with  $L_1$ , **L2** with  $L_2$ , **Ent** with  $\eta_n$  and  $\delta$ , then there exist constants  $L_0$  and  $C_0 = C(L_1, L_2)$ , such that for all  $L \geq L_0$  and all  $n \geq n_0$

$$P_{\xi^0, \beta^0} \left( \sup_{\theta \in \bar{\Theta}(L\delta)} \frac{|\Delta \tilde{C}_n(\theta) - \Delta C_n(\theta)|}{|\xi - \xi'|_{w_2}^2 + \|\beta - \beta'\|^2} \geq w_{\max} \eta_n \right) \leq \exp(-C_0 L^2 \eta_n^2 \delta^2 w_{\max}) \quad (36)$$

where  $\bar{\Theta}_n(\epsilon) = \{(\xi, \beta) : |\xi - \xi'|_{w_2}^2 + \|\beta - \beta'\|^2 \geq \epsilon^2\}$  and  $\Delta \tilde{C}_n(\theta)$ ,  $\Delta C_n(\theta)$  given in (26) and (27).

**Proof:** We will apply a small modified version of Lemma 3.4. of van de Geer (1990), with  $\Lambda = \Theta^c$  and the semimetric  $d$  given in (32) and  $Z_n(\lambda)$  from (28) with  $Z_n(\lambda^0) = 0$ . The used modified version is: Under the entropy condition on  $\Theta^c$  with respect to the metric  $d$ , that is for all  $\delta', \eta_n > 0$ , with  $\delta' \eta_n > 1$

$$\lim_{L \rightarrow \infty} \frac{\int_0^1 \sqrt{H_{\Theta^c, d}(uL\delta', L\delta' D)} du}{L\delta' \eta_n} = 0, \quad (37)$$

it holds

$$P_{\xi^0, \beta^0} \left( \sup_{d(\theta, \theta^0) > L\delta'} \frac{|Z_n(\theta)|}{d^2(\theta, \theta^0)} \geq \eta_n \right) \leq \exp(-\eta_n^2 C_0 L^2 \delta'^2). \quad (38)$$

For  $\delta' = \delta \sqrt{w_{\max}}$  from (38) follows the result. The difference to the lemma of van de Geer is that we have  $\sqrt{n} = \eta_n$ . The proof of this modification has the same steps, but we have to change  $\sqrt{n}$  to  $\eta_n$  in the entropy condition and in the exponential rate. Her assumptions on  $Z_n(\lambda)$  are not needed in the  $L_1$ -context, since the main property she used in the proof is (33), that the process is sub-Gaussian. It remains to check the entropy condition on  $\Theta^c$  (37). For Cartesian products  $A = A_1 \times A_2$  with  $a = (a_1, a_2)$  and  $d_A^2(a, a') \leq d_{A_1}^2(a_1, a'_1) + d_{A_2}^2(a_2, a'_2)$ , we know the following inequality:

$$H_{A, d_A}(\epsilon, D) \leq H_{A_1, d_{A_1}}\left(\frac{\epsilon}{2}, D\right) + H_{A_2, d_{A_2}}\left(\frac{\epsilon}{2}, D\right). \quad (39)$$

It can be derived in a similar way as the formula (7) in Lorentz (1966), on page 152. Here we have  $A = \mathcal{X}^{(n)} \times \mathcal{B}$  and  $d_{A_1}^2(a_1, a'_1) = w_{\max} |\xi - \xi'|_{w_2}^2$ , and  $d_{A_2}^2(a_2, a'_2) = w_{\max} \|\beta - \beta'\|^2$ ,

$$H_{A_1, d_{A_1}}(\epsilon \sqrt{w_{\max}}, D \sqrt{w_{\max}}) = H_{\mathcal{X}^{(n)}, \|\cdot\|_{w_2}}(\epsilon, D).$$

$\mathcal{B}$  is a set of fixed dimension  $p$ , therefore the local entropy is bounded:

$$H_{A_2, d_{A_2}}(\epsilon \sqrt{w_{\max}}, D \sqrt{w_{\max}}) = H_{\mathcal{B}, \|\cdot\|}(\epsilon, D) \leq p \ln \left( \frac{2D}{\epsilon} \sqrt{p} \right).$$

Thus it suffices to require the entropy condition (37) for  $H_{A_1, d_{A_1}}$  with  $\delta' = \delta \sqrt{w_{\max}}$  only, that is, assume **Ent**.  $\square$

## 4.2 The separation condition

The aim of this subsection is to verify the separation condition (21) for the  $L_1$ -contrast in (27). First we quote a result by Oberhofer (1982) in the form given by van de Geer (1990).

**Lemma 4** *Suppose  $\varepsilon$  r.v. whose distribution satisfies **E** with constants  $D_0$  and  $\kappa_\varepsilon$  then for all  $|\Delta| \leq \Delta_{\max}$*

$$\kappa_\varepsilon \frac{D_0}{\Delta_{\max}} |\Delta|^2 \leq E(|\varepsilon + \Delta| - |\varepsilon|) \leq |\Delta|. \quad (40)$$

**Proof:** This is the i.i.d. version of Lemma 4.2. of van de Geer (1990).  $\square$

Applying Lemma 4 to the  $L_1$ -contrast with

$$\Delta_{\max} = \max_i \max_{\xi_i} \max_{\beta \in \mathcal{B}^c} \left\{ \left| g(\xi_i^0, \beta^0) - g(\xi_i, \beta) \right| + \left| h(\xi_i^0, \beta^0) - h(\xi_i, \beta) \right| \right\},$$

we obtain

$$C_n(\theta) - C_n(\theta^0) \geq \kappa_\varepsilon \frac{D_0}{\Delta_{\max}} D_n(\xi, \beta), \quad (41)$$

with

$$D_n(\xi, \beta) = \left| g(\xi, \beta) - g(\xi^0, \beta^0) \right|_{w_1}^2 + \left| h(\xi, \beta) - h(\xi^0, \beta^0) \right|_{w_2}^2. \quad (42)$$

Now we need separation conditions on  $g$  and  $h$  also, such that it is possible to estimate

$$D_n(\xi, \beta) \geq \rho \left( d(\theta, \theta_{(n)}) \right) \quad (43)$$

for an appropriately chosen metric  $d$  on the parameter space. Deriving (43) means solving the identification problem in semiparametric models. In the same way this problem occurs also in the  $L_2$ -norm theory. This is

the main reason, why we are not able to obtain a nice consistency result for the parameter of interest  $\beta$  in the general setting (1), (2). From now on we consider the models separately. We say a regression function  $g$  fulfills the contrast condition **Con** iff

**Con**  $\exists n_0 \forall n \geq n_0 \exists a_n, 0 < a_n < \infty, \forall \xi \in \mathcal{X}^{(n)} \forall \beta, \beta' \in \Theta$

$$|g(\xi, \beta) - g(\xi', \beta)|_{w_1} \geq a_n \|\beta - \beta'\|.$$

Under **Con** for  $g$  we have in the nonlinear regression model (5), that

$$D_n(\xi, \beta) = |g(\xi^0, \beta) - g(\xi^0, \beta^0)|_{w_1}^2 \geq a_n^2 \|\beta - \beta^0\|^2. \quad (44)$$

In the nonlinear semiparametric model (13) the identification condition **ID** implies

$$2 \left( m(z, \beta) - m(z, \beta^0), f(z) - f^0(z) \right)_{w_1} = 0,$$

and under **Con** for  $m$  we have here also

$$\begin{aligned} D_n(\xi, \beta) &= |m(z, \beta) - m(z, \beta^0) + f(z) - f^0(z)|_{w_1}^2 \\ &= |m(z, \beta) - m(z, \beta^0)|_{w_1}^2 + |f(z) - f^0(z)|_{w_1}^2 \geq a_n^2 \|\beta - \beta^0\|^2. \end{aligned} \quad (45)$$

For the nonlinear functional relation model (6), (7) the following lemma helps to solve the identification problem. This result is strongly related to the Lemma 1 in Zwanzig (1990). Define

$$L_n(\xi, \beta) = |g(\xi, \beta) - g(\xi^0, \beta)|_{w_1}^2 + |g(\xi^0, \beta) - g(\xi^0, \beta^0)|_{w_1}^2 + |\xi - \xi^0|_{w_2}^2. \quad (46)$$

**Lemma 5** Under **L1**,  $\exists n_0 \exists \tau > 0 \forall n \geq n_0 \forall \xi, \xi^0 \in (\mathcal{X}^{(n)})^c \forall \beta, \beta^0 \in \mathcal{B}^c$  such that

$$|g(\xi, \beta) - g(\xi^0, \beta^0)|_{w_1}^2 + |\xi - \xi^0|_{w_2}^2 > \tau L_n(\xi, \beta). \quad (47)$$

**Proof:** Inside of this proof let us use the abbreviations  $g(\xi^0, \beta^0) = g^{00}$  and  $g(\xi^0, \beta) = g^0$ . By adding  $\pm g^0$  in  $|g - g^{00}|_{w_1}^2$ , we obtain

$$D_n(\xi, \beta) = L_n(\xi, \beta) (1 - 2\Delta_n(\xi, \beta)) \quad (48)$$

with

$$\Delta_n(\xi, \beta) = \frac{(g^0 - g^{00}, g^0 - g)_{w_1}}{|g - g^0|_{w_1}^2 + |g^{00} - g^0|_{w_1}^2 + |\xi - \xi^0|_{w_2}^2} \quad (49)$$

for  $L_n(\xi, \beta) > 0$  and  $\Delta_n(\xi, \beta) = 0$  otherwise. It remains to show, that there exists a constant  $\tau > 0$  such that

$$\sup_{\xi \in (\mathcal{X}^{(n)})^c} \sup_{\beta \in \Theta^c} \Delta_n(\xi, \beta) \leq \frac{1}{2} - \tau. \quad (50)$$

Let  $c = \frac{(\frac{1}{2} - \tau_1)^2}{L_1}$  with  $L_1$  from (30) and  $\tau_1$  such that  $0 < \tau_1 < \frac{1}{2}$ . We will distinguish two cases:

$$\text{i) } |\xi - \xi^0|_{w_2}^2 \leq c |g^0 - g^{00}|_{w_1}^2, \quad \text{ii) } |\xi - \xi^0|_{w_2}^2 \geq c |g^0 - g^{00}|_{w_1}^2.$$

i) We apply the Cauchy-Schwarz inequality and the assumption (30)

$$\Delta_n(\xi, \beta)^2 \leq \frac{|g^0 - g|_{w_1}^2 |g^0 - g^{00}|_{w_1}^2}{L_n^2} \leq \frac{L_1 |\xi - \xi^0|_{w_2}^2 |g^0 - g^{00}|_{w_1}^2}{L_n^2}$$

with  $L_n = L_n(\xi, \beta)$  given in (46). Note  $|g^0 - g^{00}|_{w_1}^2 \leq L_n$ . Under i) we have

$$\Delta_n(\xi, \beta)^2 \leq L_1 c \left( \frac{|g^0 - g^{00}|_{w_1}^2}{L_n} \right)^2 \leq L_1 c \leq \left( \frac{1}{2} - \tau_1 \right)^2,$$

and thus in case i) (50) follows. Consider case ii). Because of

$$\begin{aligned} 0 &\leq |(g^0 - g^{00}) - (g^0 - g)|_{w_1}^2 \\ &= |g^0 - g^{00}|_{w_1}^2 + |g^0 - g|_{w_1}^2 - 2(g^0 - g^{00}, g^0 - g)_{w_1}, \end{aligned}$$

we have

$$2(g^0 - g^{00}, g^0 - g)_{w_1} \leq |g^0 - g^{00}|_{w_1}^2 + |g^0 - g|_{w_1}^2.$$

Using this and (46) we obtain for  $\Delta_n = \Delta_n(\xi, \beta)$

$$2\Delta_n \leq \frac{|g^0 - g^{00}|_{w_1}^2 + |g^0 - g|_{w_1}^2}{L_n} \leq 1 - \frac{|\xi - \xi^0|_{w_2}^2}{L_n}.$$

From the assumption (30) we get

$$2\Delta_n \leq 1 - \frac{|\xi - \xi^0|_{w_2}^2}{|g^0 - g^{00}|_{w_1}^2 + (1 + L_1) |\xi - \xi^0|_{w_2}^2}.$$

For positive  $a$ , the function  $f(x) = \frac{x}{a + (1 + L_1)x}$  is increasing in  $x$ . Using ii) we have  $ca \leq x$  and  $f(x) \geq \frac{c}{1 + c + cL_1}$ . We obtain  $2\Delta_n \leq 1 - \frac{c}{1 + (1 + L_1)c}$ . For  $\tau_2$  such that  $2\tau_2 = \frac{c}{1 + (1 + L_1)c}$ ,  $0 > \tau_2 > \frac{1}{2}$  one has under ii)  $\Delta_n \leq \frac{1}{2} - \tau_2$ . We choose  $\tau = \min(\tau_1, \tau_2)$  and get (50).  $\square$

## 5 The consistency of the $L_1$ -estimators

In this section we summarize the results above and obtain the consistency of the  $L_1$ -estimators in the different submodels.

### 5.1 The nonlinear regression model

Consider the model (5). Then we have for

$$\tilde{\beta} = \arg \min_{\beta \in \mathcal{B}^c} \sum_{i=1}^n w_{1i} \left| y_i - g(\xi_i^0, \beta) \right| \quad (51)$$

the following strong consistency result. Set

$$G_{\max} = \max_i \sup_{\beta \in \mathcal{B}^c} \left| g(\xi_i^0, \beta) - g(\xi_i^0, \beta^0) \right|. \quad (52)$$

**Theorem 1** *Suppose for the error distribution  $\mathbf{E}$  with the constants  $\kappa_\epsilon$ ,  $D_0$  and suppose for the regression function  $g$  that*

$$\exists n_0 \exists L_2 \forall n \geq n_0 \exists a_n, a_n > 0 \forall \beta, \beta' \in \mathcal{B}^c$$

$$a_n^2 \|\beta - \beta'\|^2 \leq \left| g(\xi^0, \beta) - g(\xi^0, \beta') \right|_{w_1}^2 \leq L_2 \|\beta - \beta'\|^2. \quad (53)$$

Then there exists a positive constant  $C_0$  such that for all  $L > 0$  and all  $n \geq n_0$

$$P_{\xi^0, \beta^0} \left( \|\tilde{\beta} - \beta^0\| > L \right) \leq \exp \left( - \frac{C_0}{w_{\max}} \left( \frac{a_n^2 \kappa_\epsilon D_0 L}{G_{\max}} \right)^2 \right). \quad (54)$$

**Proof:** Under (53) from (44) and from (41) with (52) follows that the separation condition (21) is fulfilled with  $\rho(\|\beta - \beta^0\|) = \frac{\kappa_\epsilon D_0 a_n^2}{G_{\max}} \|\beta - \beta^0\|^2$ . Lemma 1 gives

$$P_{\xi^0, \beta^0} \left( \|\tilde{\beta} - \beta^0\| > L \right) \leq P_{\xi^0, \beta^0} \left( \sup_{\theta \in \bar{\Theta}(L)} \frac{|\Delta C_n(\theta) - \Delta \tilde{C}_n(\theta)|}{\|\beta - \beta^0\|^2} \geq \frac{\kappa_\epsilon D_0 a_n^2}{G_{\max}} \right).$$

The entropy condition **Ent** is satisfied for the one point set. Then the result (54) is a consequence of Lemma 3 with  $\eta_n = \frac{\kappa_\epsilon D_0 a_n^2}{G_{\max} w_{\max}}$ .  $\square$

Note the result is interesting only for  $\frac{a_n^2}{G_{\max}} \geq \sqrt{w_{\max}}$ .

## 5.2 The nonlinear error-in-variables model

Consider the model (6), (7). The  $L_1$ -norm estimator is defined as

$$\tilde{\beta} = \arg \min_{\beta \in \mathcal{B}^c} \min_{\xi \in \mathcal{X}^{(n)c}} \sum_{i=1}^n w_{1i} |y_i - g(\xi_i, \beta)| + w_{2i} |x_i - \xi_i|. \quad (55)$$

Set

$$G_{\max} = \max_i \sup_{\xi \in \mathcal{X}^{(n)c}} \sup_{\beta \in \mathcal{B}^c} |g(\xi_i, \beta) - g(\xi_i^0, \beta^0)| + |\xi_i - \xi_i^0|. \quad (56)$$

Then we have the following exponential probability inequality.

**Theorem 2** Suppose for the error distribution  $\mathbf{E}$  with the constants  $\kappa_\epsilon$ ,  $D_0$  and suppose for the nuisance parameter set the entropy condition **Ent** is satisfied with  $\eta_n = \frac{a_n^2 \kappa_\epsilon D_0 \tau}{w_{\max} G_{\max}}$  and  $\delta_n$ .

Suppose for the regression function  $g$  that

$$\exists n_0 \exists L_1, L_2 < \infty, \forall n \geq n_0 \exists a_n, a_n > 0 \quad \forall \beta, \beta' \in \mathcal{B}^c \quad \forall \xi, \xi' \in \mathcal{X}^{(n)c} \quad (57)$$

$$a_n^2 \|\beta - \beta'\|^2 \leq |g(\xi, \beta) - g(\xi, \beta')|_{w_1}^2 \leq L_2 \|\beta - \beta'\|^2 \quad (58)$$

$$\text{and} \quad |g(\xi, \beta) - g(\xi', \beta)|_{w_1}^2 \leq L_1 |\xi - \xi'|_{w_2}^2.$$

Then there exist positive constants  $C_0, L_0$  such that for all  $L > L_0$  and all  $n \geq n_0$

$$P_{\xi^0, \beta^0} \left( \|\tilde{\beta} - \beta^0\| > L\delta_n \right) \leq \exp \left( -\frac{\delta_n^2}{w_{\max}} \left( \frac{\kappa_\epsilon D_0 a_n^2}{G_{\max}} \right)^2 C_0 L^2 \right). \quad (59)$$

**Proof:** Without loss of generality we set  $a_n < 1$ . Under (56) from Lemma 4 and under (58) from Lemma 5 follows that the separation condition (21) is fulfilled with

$$\rho \left( \sqrt{\|\beta - \beta^0\|^2 + |\xi - \xi^0|_{w_2}^2} \right) = \kappa_\epsilon D_0 \frac{a_n^2 \tau}{G_{\max}} \left( \|\beta - \beta^0\|^2 + |\xi - \xi^0|_{w_2}^2 \right).$$

Lemma 1 gives

$$\begin{aligned} P_{\xi^0, \beta^0} \left( \|\tilde{\beta} - \beta^0\| > L\delta_n \right) &\leq P_{\xi^0, \beta^0} \left( \|\tilde{\beta} - \beta^0\|^2 + |\tilde{\xi} - \xi^0|_{w_2}^2 > L^2 \delta_n^2 \right) \\ &\leq P_{\xi^0, \beta^0} \left( \sup_{\theta \in \bar{\Theta}(L\delta_n)} \frac{|\Delta C_n(\theta) - \Delta \tilde{C}_n(\theta)|}{\frac{\kappa_\epsilon D_0 a_n^2 \tau}{G_{\max}} \left( \|\beta - \beta^0\|^2 + |\xi - \xi^0|_{w_2}^2 \right)} \geq 1 \right). \end{aligned}$$



The entropy condition **Ent** is assumed above explicitly. Hence the result (59) is a consequence of Lemma 3 with  $\eta_n = \frac{\kappa_\epsilon D_\sigma \tau a_n^2}{w_{\max} G_{\max}}$ .  $\square$

Consider now special cases and the unweighted case  $w_{\max} = n^{-1}$ .

**Corollary 1** *Suppose for the error distribution **E** and suppose for the regression function  $g$  that (57), (58) is valid with*

$$\frac{G_{\max}}{a_n^2} \leq \text{const}, \text{ for all } n. \quad (60)$$

Suppose i)  $\mathcal{X}^{(n)}$  defined in (9) or ii)  $\mathcal{X}^{(n)}$  defined in (10). Then

$$\tilde{\beta} \rightarrow \beta^0 \quad P_{\xi^0, \beta^0} - a.s.. \quad (61)$$

**Proof:** Under (60) in the unweighted case we have  $\eta_n \sqrt{w_{\max}} \leq \text{const} \sqrt{n}$ .

i) For  $\mathcal{X}^{(n)}$  defined in (8), the entropy is  $H_\xi(\delta, D) \leq \text{const} \frac{1}{\delta} \ln^+ \left( \frac{1}{\delta} \right)$  and the entropy condition **Ent** is satisfied with  $\delta = n^{-\frac{1}{3}} (\ln n)^{\frac{1}{3}}$ , (see Example 2.1 of van de Geer (1990)). From Theorem 2 follows that there exists a  $d$ ,  $0 < d < 1$ , for all  $\epsilon > 0$

$$\sum_{n=1}^{\infty} P_{\xi^0, \beta^0} \left( \|\tilde{\beta} - \beta^0\| > \epsilon \right) \leq \sum_{n=1}^{\infty} \exp \left( -n^d \epsilon \text{const} \right) \quad (62)$$

$$< \text{const}(n_0) \sum_{n=n_0}^{\infty} n^{-2} < \infty. \quad (63)$$

We obtain the statement by the Lemma of Borel Cantelli.

ii) For  $\mathcal{X}^{(n)}$  defined in (10) the entropy can be derived from the classical result of Kolmogorov and Tichomirov (1960), for the sup-norm  $|f - f^0|_{\sup} = \max_{x \in [0,1]} |f(x) - f^0(x)|$

$$H_{\mathcal{X}^{(n)}, |\cdot|_{\sup}}(\delta, D) \leq \text{const} \left( \frac{C}{\delta} \right)^{\frac{1}{m+\alpha}}.$$

Since under the design assumption **D**  $\max_i |z_i - z_{i-1}| \leq c_d$  for  $n \geq n_0$  and since

$$\begin{aligned} |\xi - \xi^0|_{w_2} &\leq \max_i |f(z_i) - f^0(z_i)| \leq |f - f^0|_{\sup} \\ &\leq \max_i |f(z_i) - f^0(z_i)| + 2Lc_d \end{aligned}$$

we have  $\mathcal{X}^{(n)} \subseteq \{f : |f - f^0|_{\sup} \leq 3(C + L)\}$ . Thus

$$H_\xi(\delta, D) \leq \text{const} \left( \frac{1}{\delta} \right)^{\frac{1}{m+\alpha}}. \quad (64)$$

Then the entropy condition **Ent** is fulfilled for  $\delta = n^{-b}$  with  $b = \frac{m+\alpha}{2(m+\alpha)+1} > 0$  and the result (61) follows by the same arguments as in (62).  $\square$

### 5.3 The nonlinear semiparametric model

Consider the model (13) with design condition **D** and identification condition **ID**. Then we can derive on the same way as above the strong consistency of the  $L_1$ -estimator

$$\tilde{\beta} = \arg \min_{\beta \in \mathcal{B}^c} \min_{f \in \mathcal{K}_{m,\alpha}(C,L)} \sum_{i=1}^n w_i |y_i - m(z_i, \beta) - f(z_i)|.$$

**Theorem 3** Suppose for the error distribution **E** and suppose for the function  $m$  that

$$\begin{aligned} & \exists n_0 \exists L_2 < \infty, \exists a > 0 \forall n \geq n_0 \quad \forall \beta, \beta' \in \mathcal{B}^c \\ & a^2 \|\beta - \beta'\|^2 \leq |m(z, \beta) - m(z, \beta')|_{w_1}^2 \leq L_2 \|\beta - \beta'\|^2 \end{aligned} \quad (65)$$

and

$$G_{\max} = \max_i \sup_{\beta \in \mathcal{B}^c} |m(z_i, \beta) - m(z_i, \beta^0)| \leq \text{const}. \quad (66)$$

Then

$$\tilde{\beta} \rightarrow \beta^0 \quad P_{\xi^0, \beta^0} - a.s. \quad (67)$$

**Proof:** Under **ID** and (65) from (41) and (45) the separation condition (21) of Lemma 1 is satisfied with  $\Delta_{\max} = G_{\max} + 2C$ , where  $G_{\max}$  from (66) and where  $C$  from (11),

$$\rho(\|\beta - \beta^0\|) = \kappa_\epsilon D_o \frac{a^2}{G_{\max} + 2C} \|\beta - \beta^0\|^2.$$

Because of (64) for  $\eta_n \sqrt{w_{\max}} \leq \text{const} \sqrt{n}$  and for  $\delta = n^{-b}$  with  $b = \frac{m+\alpha}{2(m+\alpha)+1} > 0$  the entropy condition **Ent** of Lemma 3 is valid. Then from both lemmata follows an inequality of type (62) and we obtain (67) by the same arguments as in (62).  $\square$

## References

- [1] L. Birge and P. Massart (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields* **77**, 115–150.
- [2] L. Birge and P. Massart (1994). Minimum contrast estimators on sieves. Technical Report 94, Universite de Paris-Sud.
- [3] S. van de Geer (1995). The method of sieves and minimum contrast estimators. *Math. Meth, Statist.* **4**, 20–38.
- [4] S. van de Geer (1990). Estimating a regression function. *Ann. Statist.* **18**, 907–924.

- [5] A. N. Kolmogoroff and W. M. Tichomirow (1960). *Arbeiten zur Informationstheorie III*. Mathematische Forschungsberichte, Verlag der Wissenschaften, Berlin.
- [6] A. Kukush and S. Zwanzig (1996). On an alternative estimator in nonlinear functional relations. Unpublished manuscript.
- [7] A. Kukush and S. Zwanzig (1996). On inconsistency of the least squares estimator in nonlinear functional error-in-variables models. Technical Report 12, Institute of Mathematical Stochastics, University of Hamburg.
- [8] F. Liese and Vajda I (1994). Consistency of M-estimates in general regression models. *J. Multiv. Anal.* **50**, 93–114.
- [9] F. Liese and Vajda I. (1995). Necessary and sufficient conditions for consistency of generalized M-estimates. *Metrika*, **42**, 291–324.
- [10] O. Linton (1995). Second order approximation in the partially linear regression model. *Econometrica* **63**, 1079–1112.
- [11] G. G. Lorentz (1966). *Approximation of Functions*. Holt, Rinehart and Winston.
- [12] W. Oberhofer (1982). The consistency of nonlinear regression minimizing the L1-norm. *Ann. Statist.* **10**, 316–319.
- [13] D. Pollard (1984). *Convergence of Stochastic Processes*. Springer Series in Statistics.
- [14] G. D. Richardson and B. B. Bhattacharyya (1987). Consistent L1-estimators in nonlinear regression for a noncompact parameter space. *Sankhyā A* **49**, 377–387.
- [15] A. W. van der Vaart and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.
- [16] S. Zwanzig (1980). The choice of approximative models in nonlinear regression. *Statistics* **11**, 23–47.
- [17] S. Zwanzig (1990). On consistency in nonlinear functional relations. P-MATH P-Math 10-90, Institute of Mathematics, Academy of Science GDR.

# A comparison of two LP solvers and a new IRLS algorithm for $L_1$ estimation

C.J. Adcock

*University of Westminster, U.K.*

N. Meade

*Management School Imperial College, U.K.*

*Abstract:* This paper is a comparison of two methods for computing  $L_1$  estimates of the parameter vector  $\beta$  in the linear model. The main methods in the comparison are in two groups: special purpose linear programming (LP) methods which exploit the structure of the objective function and iteratively re-weighted least squares (IRLS). The special purpose LP methods included in the review are: (i) the Barrodale and Roberts (BR) algorithm and (ii) the modified form due to Bloomfield and Steiger (BS). The IRLS methods is a new development which exploits the piecewise differentiability of the objective function and which avoids the difficulties previously associated with least squares based schemes. All algorithms have been implemented in a common language, in order to provide a better basis for comparison. To summarise: we found that our implementations of the BR & BS algorithms are generally quicker than existing implementations and general purpose LP solvers; the new IRLS algorithm is faster in circumstances where the number of observations is very large relative to the number of parameters to be estimated.

*Key words:* Regression, linear model, minimum absolute deviations, LP solvers, iteratively re-weighted least squares, piecewise differentiability.

AMS subject classification: 65J05, 62G05.

## 1 Introduction

The method of minimum absolute deviation (MAD) or  $L_1$  estimation, to give it one of the many names by which the technique is known, is a robust

method for estimating the parameters in the linear model:

$$y = X\beta + \varepsilon.$$

The objective function to be minimised is:

$$f(\beta) = \sum_{i=1}^n |y_i - \beta^T x_i| \quad (1)$$

where  $\beta \in \mathcal{R}^k$ . Minimising  $f(\beta)$  will also give maximum likelihood estimators of  $\beta$  when the  $\{y_i\}$  are random sample from the double sided exponential distribution. In this paper it is assumed that the function is not degenerate, in which case it will possess a unique minimiser,  $\beta^*$  say, at which up to  $k$  of the residuals:

$$r_i = y_i - \beta^{*T} x_i \quad (2)$$

will satisfy:

$$r_i = 0, \quad i \in B, \text{ say} \quad (3)$$

The set  $B$  defines a set of basis vectors  $x_i$  which span  $\mathcal{R}^k$ . If the subsets of  $\{y_i\}$  and  $\{x_i\}$  defined by  $B$  are denoted by  $y^*$  and  $X^*$ , then  $\beta^*$  satisfies:

$$X^* \beta^* = y^* \quad (4)$$

There are two general types of algorithm for calculating MAD estimates of the parameters of a linear model. The first type relies on the fact that the objective function  $f(\beta)$  at equation (1) can be formulated as a linear program, Charnes et al (1955). This type of method includes a procedure due originally to Barrodale and Roberts (1973) [henceforth BR] which exploits the fact that the MAD objective function may be written as an LP with special structure. Other LP methods, which also exploit the special structure, have been reported by Bloomfield & Steiger (1980) and Seneta & Steiger (1984). The second type of procedure uses iteratively re-weighted least squares (IRLS). This method was reported by Schlossmacher (1973) and Fair (1974). According to Bloomfield and Steiger (1984, page 259) [henceforth BS], however, it was due originally to Beaton & Tukey (1974), Comparative studies are reported in BS. There is a comprehensive review of algorithms in Dielman (1992).

One of the motivations for this paper is that, although there are very strong similarities between the special purpose LP solvers mentioned above, comparisons are limited by the fact to date, and to the best of our knowledge, the software implementations are quite distinct. Self evidently, comparison is greatly facilitated if the software is written in the same language

and meets similar design criteria. A second motivation is our wish to develop algorithms to minimise mixed objective functions of the form:

$$f(\beta) = \alpha \sum_{i=1}^n (y_i - \beta^T x_i)^2 + (1 - \alpha) \sum_{j=1}^m |Y_j - \beta^T X_j|. \quad (5)$$

These objective functions arise in other robust methods and in dynamic estimation schemes in which the current parameter estimates are (approximately or asymptotically) normally distributed. Robust estimation methods using similar convex objective functions, in which there is a modulus term, have been studied by Dodge & Jureckova (1991 and 1992) and related computational aspects are reported in Dodge et al (1991). Objective functions of the form in (5) also arise in portfolio optimisation when the conventional quadratic programming formulation is extended by the inclusion of transactions costs. In the case of portfolio optimisation, the minimisation of  $f(\beta)$  is invariably carried out subject to a number of linear inequality constraints of the values of the parameter vector  $\beta$ . Algorithms to minimise  $f(\beta)$  given at equation (5) may use IRLS methods - see Adcock & Meade (1995) for an example of IRLS used in portfolio optimisation. However, the well reported deficiencies of IRLS have prompted us to consider the general question of algorithms for MAD estimation ab initio.

The purpose of this paper is therefore to compare the solutions times of the main established special purpose LP and IRLS algorithms for MAD estimation. We use new implementations of the Barrodale and Roberts and the Bloomfield and Steiger algorithms. We also present a new procedure for IRLS. The algorithm that we describe in Section 3 of this paper is different from the scheme developed by Schlossmacher and others in that it exploits the piecewise differentiability of the objective function. All algorithms included in the comparison have been re-implemented in a single programming language and in a similar programming style. The aim is to provide a fair basis for comparison of the solution times. In addition, and as reported below, we have found that our new code offers performance improvements over existing software.

The structure of the paper is as follows. Section 2 describes methods based on special purpose linear programming methods. Section 3 presents our scheme which uses iteratively re-weighted least squares. We compare performance using a number of different data sets in Section 4. The final section of the paper contains a summary and concluding remarks.

## 2 Linear programming methods

Following Charnes et al (1955), the objective function at (1) may be written

exactly as:

$$f(\beta) = \sum_{i=1}^n |y_i - \beta^T x_i| = \sum_{i=1}^n |e_i^+ + e_i^-| \quad (6)$$

where:

$$e_i^+, e_i^- \geq 0 \quad (7)$$

and where the corresponding  $n$ -vectors  $e^+$  and  $e^-$  satisfy:

$$y - X\beta = e^+ - e^- \quad (8)$$

Minimisation of  $f(\beta)$  is now a linear programming problem involving the  $2n+p$  variables  $e^+$ ,  $e^-$  and  $\beta$ , together with the  $n$  equality constraints at (8) and the  $2n$  non-negativity constraints at (7). Calculation of the minimiser  $\beta^*$  may be undertaken using standard LP solvers. However, these methods are very slow when compared with special purpose solvers which can exploit the structure of the LP formulation of the objective function. The following results are summarised from BS who provide further details and proof of the properties of the algorithms.

## 2.1 The Barrodale and Roberts (BR) algorithm

The BR algorithm may be viewed as row and column operations on a  $(n+k) \times (k+1)$  matrix  $A$ . The initial value of  $A$  is:

$$A = \begin{bmatrix} X & y \\ I & 0 \end{bmatrix}$$

In the steps below the elements of  $A$  are  $\{a_{ji}\}$ . Note that in this notation  $i$  indexes columns and  $j$  rows rather than the more conventional arrangement.

**Step 1** compute:

$$\begin{aligned} g_i &= \sum_j |a_{ji}|, \quad j \text{ over } a_{jk+1} = 0 \\ h_i &= \sum_j a_{ji} \cdot \text{sign}(a_{jk+1}) \\ l_i &= \min(g_i - h_i, g_i + h_i) \end{aligned}$$

where:  $i = 1(1)k$  and  $j = 1(1)n$ .

**Step 2** determine:

$$p = I \text{ where } l_I = \min_i(l_i)$$

If  $l_I > 0$  then go to **Step 5**

**Step 3** determine the pivot row  $q$  by finding the MAD estimate of  $t$  in:

$$f = \sum_j |a_{jk+1} - ta_{jp}|$$

ie, find  $t^* = y_q/x_{qp}$ . It should be noted that this requires a sort routine and that suitable choice of sorter can affect algorithm timings.

**Step 4** pivot on row  $q$  column  $p$ , ie compute new columns  $a'_{.j}$  of  $A$ :

$$\begin{aligned} a'_{.p} &= a_{.p}/a_{qp} \\ a'_{.j} &= a_{.j} - a_{qj}a_{.p} \quad j \neq p \end{aligned}$$

and go to **Step 1**.

**Step 5** The minimiser  $\beta^*$  may be recovered from the block of  $A$  corresponding to the initial zero vector, ie in column  $k+1$  rows  $n+1$  through  $n+k$ , together with a sign reversal. Specifically:  $\beta_i^* = -a_{n+i,k+i}$ .

## 2.2 Bloomfield and Steiger (BS) algorithm

There are two modifications to the BR algorithm which are described in BS. One is due to Bloomfield & Steiger (1980) themselves. It is the essentially the same as BR except that in Step 1 a heuristic is used to compute  $g_i$  and  $h_i$ . In an obvious notation:

$$g_i^{BS} = g_i^{BR} / \sum_j |a_{ji}|, \quad h_i^{BS} = h_i^{BR} / \sum_j |a_{ji}|.$$

According to Bloomfield & Steiger, the BS algorithm often converges more quickly than the original BR procedure.

## 3 Algorithms based on iteratively reweighted least squares (IRLS)

Iteratively re-weighted least squares (IRLS) is an alternative approach to LP. The usual approach is to write  $f(\beta)$  identically as:

$$f(\beta) = \sum_{i=1}^n (y_i - \beta^T x_i)^2 / |y_i - \beta^T x_i| \quad (9)$$

If  $\beta_p$  is an approximation to  $\beta^*$ , a new approximation is computed by minimising the sum of squares:

$$f(\beta) = \sum_{i=1}^n (y_i - \beta^T x_i)^2 / |y_i - \beta_p^T x_i| \quad (10)$$



which is equivalent to differentiating (9) while holding the  $|y_i - \beta^T x_i|$  terms in the denominator fixed. In conventional OLS matrix notation, the equation for the new approximation  $\beta_{p+1}$  may be written as:

$$X^T W_p X \beta_{p+1} = X^T W_p y \quad (11)$$

where:

$$W_p = \text{Diag}\{1/|y_i - \beta_p^T x_i|\} \quad (12)$$

Since it is known that, at the minimiser  $\beta^*$ , up to  $k$  of the residuals  $y_i - \beta^{*T} x_i$  will equal zero, this procedure requires some modification or it will fail as the elements of  $W_p$  become very large. A common modification is to define:

$$W_p = \text{Diag}\{W_{pi}\}$$

where:

$$\begin{aligned} W_{pi} &= 0 \text{ if } y_i - \beta_p^T x_i = 0 \\ &= 1/|y_i - \beta_p^T x_i| \text{ otherwise} \end{aligned} \quad (13)$$

This algorithm does not run without problems in practice and it is criticised in BS on the grounds that it is slow and prone to be unstable. A modification to the basic IRLS scheme was introduced by Adcock & Meade (1995). They note that at the points at which it exists, the vector of partial derivatives of  $f(\beta)$  is:

$$f'(\beta) = 2X^T W X - 2X^T W y - \left(\sum_A x_i - \sum_B x_i\right) \quad (14)$$

where:

$$W = \text{Diag}\{1/|y_i - \beta^T x_i|\} = \text{Diag}\{W_i\} \quad (15)$$

$A = \{i : (y_i - \beta^T x_i) < 0\}$  and  $B = \{i : (y_i - \beta^T x_i) > 0\}$ . This suggests the iterative scheme with limiting equation:

$$\beta = (X^T W X)^{-1} \{X^T W y + 0.5(\sum_A x_i - \sum_B x_i)\} \quad (16)$$

which may be re-arranged as:

$$\begin{aligned} \beta_{p+1} &= \beta_p - 0.5(X^T W_p X)^{-1}(\sum_A x_i - \sum_B x_i) \\ &= \beta_p - 0.5\delta_p \text{ say,} \end{aligned} \quad (17)$$

where  $\delta_p$  is the step length at iteration  $p$ . For practical purposes, the modification of  $W_p$  described at (13) is employed. The process terminates

when the absolute change in the value of the objective function is less than a given tolerance and the absolute change in each estimated parameter value is also less than a set tolerance.

To improve convergence, we also employ a number of empirical modifications of the scheme. When the process described above converges, to  $\beta'$  say, the label set corresponding to the  $k$  smallest absolute values of the residuals  $|y_i - \beta'^T x_i|$  is used to define a basis  $B^*$ . If the subsets of  $\{y_i\}$  and  $\{x_i\}$  defined by  $B^*$  are collectively denoted by  $y^*$  and  $X^*$ , then the minimiser  $\beta^*$  is computed as the solution to:

$$X^* \beta^* = y^* \quad (18)$$

as long as  $f(\beta^*) < f(\beta')$ . Otherwise the minimiser is taken as  $\beta'$ . It should be noted that if  $f(\beta^*) \geq f(\beta')$  then  $\beta^*$  cannot be the minimiser which implies that  $\beta'$  is not the minimiser either. However, for the data sets described below, this algorithm always converged to the correct solution. That is, the solution computed by the IRLS method described above was always the same as that computed by the BR or BS algorithms. In this context the same is taken to mean an accuracy equal or better than the process termination parameters.

## 4 A comparison of performance

In order to compare the computational efficiency of the MAD algorithms described, the times taken by the algorithms to solve a range of problems were measured. If a general data set is denoted as  $\{y_i; x_{ij}; i = 1(1)n, j = 1(1)k\}$  then the test data was generated by the following procedure.

1. The  $x_{i1}$  are sampled from a uniform distribution on the interval  $(-1000, 1000)$ , for  $i = 1(1)n$ .

2. The remaining  $x_{ij}$ ,  $j = 2(1)k$  are generated by the equation:

$$x_{ij} = u_{ij} + c_j x_{1j}$$

where the  $u_{ij} \sim U(-1000, 1000)$  for  $i = 1(1)n$ ,  $j = 2(1)k$  and where the constants  $c_j$  were set to 2.0,  $j = 2(1)k$ , to control the collinearity between the  $x_{ij}$ .

3. The dependent variables were generated by the equation:

$$y_i = \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i$$

where the true coefficients  $\beta_j \sim U(-1, 1)$ .

MAD estimation corresponds with the maximum likelihood estimation of the parameters  $\{\beta_j\}$  where the residuals  $\varepsilon_i$  follow a double sided exponential distribution. BS use both the Gaussian and the Pareto in their algorithm comparisons. This led us to generate three different data sets where the values of the error term  $\varepsilon_i$  are sampled from three different distributions, namely the double sided exponential, the Gaussian and the Pareto with appropriate parametrisation. In order to standardise the error distributions, the interquartile range for each distribution was set to  $(-50, 50)$ . Data sets were generated with the combination of numbers of observations and numbers of variables shown in Table 1. For each error distribution, four sets of observations for each combination marked with tick, giving  $480 = 3 * 4 * 40$  data sets in all.

Table 1: Data sets generated

|                     | Number of Observations |    |    |    |     |     |     |      |      |      |       |
|---------------------|------------------------|----|----|----|-----|-----|-----|------|------|------|-------|
|                     |                        | 10 | 20 | 50 | 100 | 200 | 500 | 1000 | 2000 | 5000 | 10000 |
| Number of variables | 1                      | *  | *  | *  | *   | *   | *   | *    | *    | *    | *     |
|                     | 2                      | *  | *  | *  | *   | *   | *   | *    | *    | *    | *     |
|                     | 5                      | *  | *  | *  | *   | *   | *   | *    | *    |      |       |
|                     | 10                     |    | *  | *  | *   | *   | *   | *    | *    |      |       |
|                     | 20                     |    |    | *  | *   | *   | *   | *    |      |      |       |

The BR algorithm and the iteratively re-weighted least squares (IRLS) algorithm were used to estimate the parameters of each data set. Both algorithms were programmed in Fortran 77. We implemented a new version of the BR algorithm. This follows the procedure described in Section 2. The computations were performed on a Silicon Graphics workstation. Since the solution times for small problems was very short, the actual time measured was that to solve the same problem ten times.

The convergence parameters in the IRLS algorithm were set so that the algorithm was deemed to have converged if:

$$f(\beta_{p-1})/f(\beta_p) - 1 \leq 10^{-6} \text{ and } \frac{1}{k} \sum_{j=1}^k |\beta_{j,p} - \beta_{j,p-1}| \leq 10^{-6} .$$

It is well known that the values of these tolerances can affect solution time substantially. The above values were chosen after some initial investigation with the aim of ensuring that, for those data sets where the new IRLS method converged satisfactorily, the numerical discrepancies between IRLS and LP solutions were small. As already reported in Section 3, for the data sets considered, all three method always converged to the same solution.

In order to gain some understanding from these timings, a model of the computation time,  $T$  say, as a function of the number of variables,  $k$  say, and the number of observations,  $n$  say, was constructed. Anderson & Steiger (1982) proposed a model of the form:

$$T = \gamma_0 + \gamma_1 n + \gamma_2 k + \gamma_3 nk + \eta$$

Trials with this model were unsatisfactory in that the estimated values of  $T$  were negative for small values of  $k$ . Beasley (1990) suggested a log-linear model for the timing of the LP solutions. This provided a basis for the following model:

$$\ln(T) = \gamma_0 + \gamma_1 \ln(k) + \gamma_2 \ln(n) + \gamma_3 \ln(\ln(n)) + \eta'$$

The coefficients  $\eta$  were estimated by minimising  $\sum |\eta'|$ . The estimated coefficients for the BR algorithm and for IRLS are in Table 2 which shows the estimated coefficients over all data sets and for each error distribution separately.

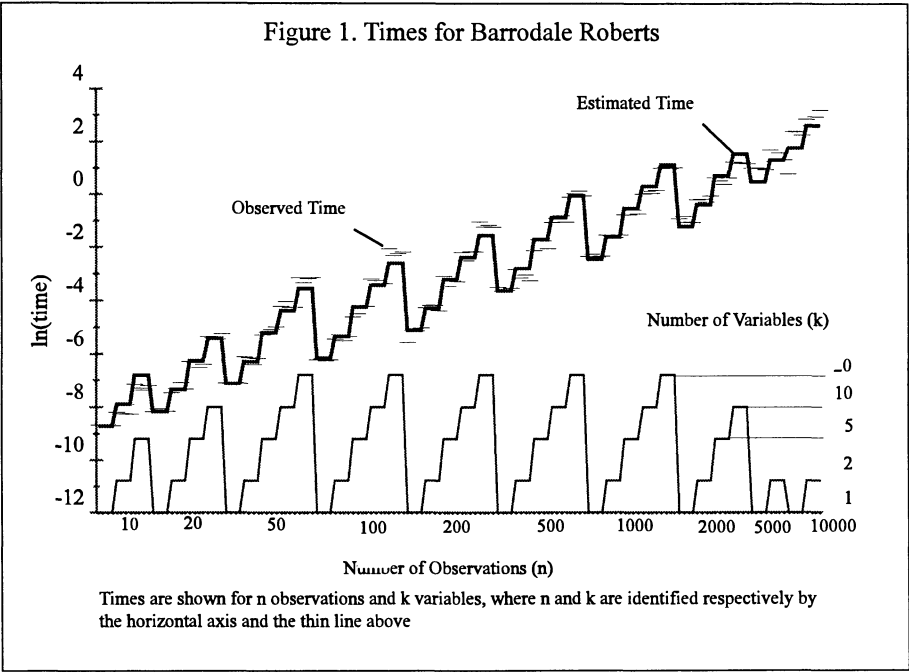
Table 2: Estimates of parameters in timing model

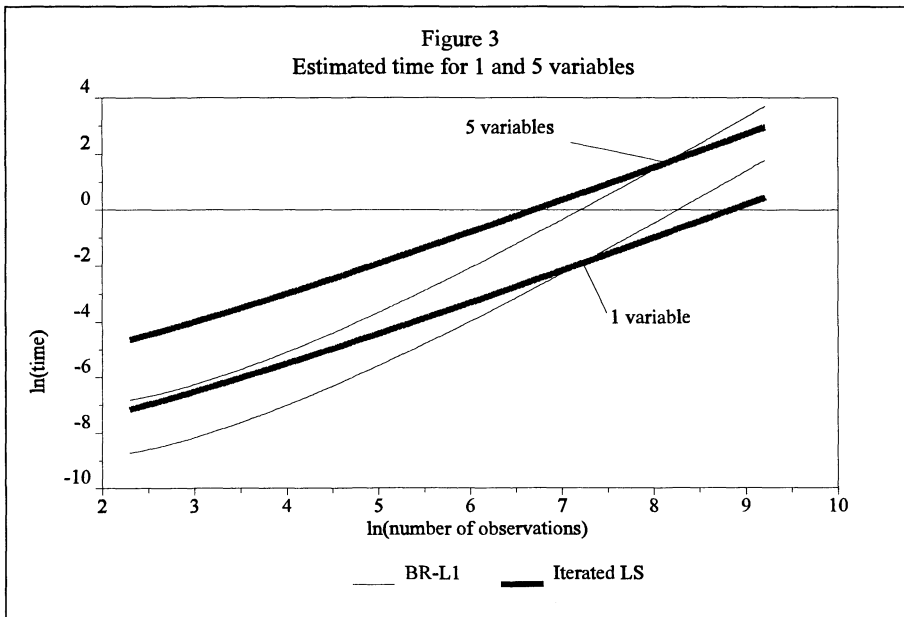
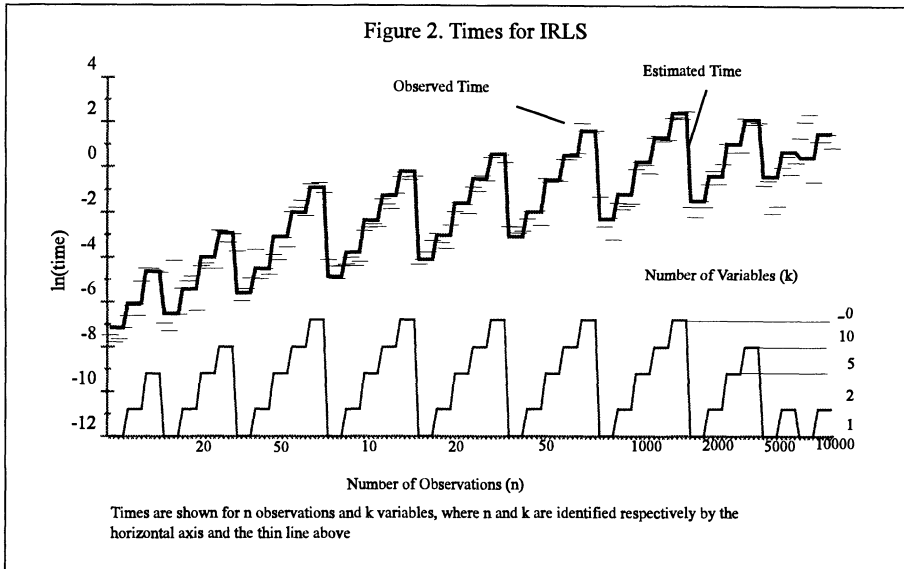
| Error Distribution |             | Barrodale and Roberts | Iterated Least Squares |
|--------------------|-------------|-----------------------|------------------------|
| Neg. Exp.          | 0           | -10.76                | -9.57                  |
|                    | 1           | 1.19                  | 1.58                   |
|                    | 2           | 2.32                  | 1.29                   |
|                    | 3           | -3.96                 | -0.89                  |
|                    | mean $ e' $ | 0.20                  | 0.38                   |
|                    |             |                       |                        |
| Gaussian           | 0           | -10.68                | -9.51                  |
|                    | 1           | 1.18                  | 1.58                   |
|                    | 2           | 2.38                  | 1.17                   |
|                    | 3           | -4.21                 | -0.55                  |
|                    | mean $ e' $ | 0.19                  | 0.35                   |
|                    |             |                       |                        |
| Pareto             | 0           | -10.73                | -8.59                  |
|                    | 1           | 1.19                  | 1.74                   |
|                    | 2           | 2.35                  | 1.54                   |
|                    | 3           | -4.15                 | -2.15                  |
|                    | mean $ e' $ | 0.19                  | 0.31                   |
|                    |             |                       |                        |
| Combined           | 0           | -10.70                | -9.30                  |
|                    | 1           | 1.19                  | 1.56                   |
|                    | 2           | 2.34                  | 1.30                   |
|                    | 3           | -4.08                 | -1.02                  |
|                    | mean $ e' $ | 0.20                  | 0.42                   |
|                    |             |                       |                        |

Comparison of the parameters for the time taken by the BR algorithm shows a consistency between the different error distributions. There is no discernible difference in timings caused by the choice of error distribution. The timings for IRLs are more dispersed about the fitted equation, showing that for the same dimension of problem there is twice as much variability in solution time in terms of mean absolute error than for the BR algorithm. For a problem of the same dimensions, the Pareto distribution appears to require a greater solution time than the other distributions by a factor of about two. Anderson & Steiger (1982) conjectured that the timing of the BR algorithm increased with approximately  $n^2$ . Their conjecture is confirmed by these results which suggest that timings are proportional to  $n^{2.34}$ . The term in  $\ln(n)$  adjusts this slightly over a wide range of numbers of observations. The timing increases slightly faster than linearly in terms of the number of variables: it is proportional to  $k^{1.2}$ .

The reliance of the IRLS technique on matrix inversion means that the computation time increases faster than linearly with  $k$ . In fact it appears that computation time is proportional to  $k^{1.5}$ . However, in contrast to the BR algorithm, the IRLS computation time increases only a little faster than linearly in the number of observations, ie it is proportional to  $n^{1.3}$ .

The actual times plotted along with the estimated times for the BR and IRLS algorithms are shown in Figures 1, 2 and 3.





The greater variability of the IRLS solution times is apparent. More interestingly, the superiority of the BR algorithm decreases as the number of observations increases. This crossover is clearly brought out in Figure 4 where the expected solution times for each algorithm is shown for one and five variables. Thus, for one variable and about 1500 observations, the expected solution times are the same. For a larger number of observations,

one would expect IRLS to be quicker. For five variables, the cross over occurs at about 3300 observations. We note the results of a simulation study of times for MAD estimation of simple linear regression reported by Armstrong & Frome (1976), who also compared IRLS and LP methods. Although their results cannot be compared directly with ours, they also found a similar superiority of LP over IRLS as far as mean solution time is concerned. However, as the table in their paper indicates, the superiority declines as the number of observations increases.

As indicated above, as part of this study we wrote up a fresh modular implementation of the BR algorithm. This has a built in facility to be converted to the Bloomfield and Steiger algorithm which we also used to compute solution times for the above data sets. The parameter estimates in the above timing model for the BS algorithm are shown in Table 3.

As the table shows, there is no discernible difference between the two solution times for the BR and BS algorithms. The estimated value of  $\gamma_2$  for the BS algorithm is broadly consistent with Bloomfield and Steiger's contention that the computation time is linear in the number of observations  $n$ . However, the more interesting coefficient is that for the number of variables which shows that the solution time increases less than linearly. It is in fact proportional to  $k^{0.67}$ . The disadvantage with this implementation of the BS algorithm is that it is slow for small problems.

Table 3: Estimates of parameters in timing model for BR/S algorithm

| Error Distribution | BR/S        |       |
|--------------------|-------------|-------|
| Combined           | 0           | -2.07 |
|                    | 1           | 0.67  |
|                    | 2           | 1.21  |
|                    | 3           | -4.11 |
|                    | mean $ e' $ | 0.44  |

Finally, we also used two other MAD algorithms which are in the public domain and computed the solution times. The first was the version of the BR algorithm provided in the NAG library. The second was a line by line re-implementation of the BR algorithm as described in Barrodale and Roberts (1974). We found that, with respect to the number of variables, these two algorithms are similar to our modular implementation of BR and BS in that the solution times vary linearly with the number of variables. However, we also found that the solution time of these public domain algorithms increases much more steeply with the number of observations.

It should be noted that none of the above data sets are degenerate and that, because of the use of the uniform distribution, there are no extreme outliers in the independent  $x$  variables. Investigation of conditions in which

the new IRLS algorithm, or indeed BR/BS, might fail remains a topic for further investigation. At present, we are inclined to the view that some algorithm problems may be due in part to the use of an inappropriate programming language.

## 5 Summary and concluding remarks

In this paper, we have presented and compared two methods for MAD estimation of the parameters in a linear model. They are described in detail, with accompanying pseudo-code in Adcock & Meade (1997). There is no single algorithm that is superior to all the others, at least as far as the data sets that we have investigated are concerned. The IRLS algorithm that we have described in this paper did not suffer from any problems of convergence or of numerical accuracy. Furthermore, we found it to be faster than all of the special purpose LP algorithms for data sets where the number system is greatly over-determined. We did not find significant differences between the BR and BS algorithms. Our investigations of various implementations of the BR algorithm indicated that the speed of the sort routine, which is a necessary step in each iteration of the algorithm, is crucial to the overall time taken. According to Dielman (1992), it is also likely that the algorithm due to Armstrong et al (1979), which employs an LU decomposition of the basis matrix, may lead to further performance improvements. This remains a topic for further investigation.

## References

- [1] Adcock C. J. and Meade N. (1995). A simple algorithm to incorporate transactions costs in quadratic optimisation. *European Journal of Operational Research* **79**, 85-94..
- [2] Adcock C. J. and Meade N. (1997). Two algorithms for minimum absolute deviation estimation in the linear model. In preparation.
- [3] Anderson D. and Steiger W. L. (1982). A comparison of methods for discrete L1 curve fitting. Technical Report 69, Department of Computer Science, Rutgers University.
- [4] Armstrong R. D. and Frome E. L. (1976). A comparison of two algorithms for absolute deviations curve fitting. *J Am. Statist. Assoc.* **71**, 328-330.
- [5] Armstrong R. D., Frome E. L. and Kung D. S. (1979). A revised simplex algorithm for the absolute deviation curve-fitting problem. *Commun. Statist. - Simul. Comp.* **B 8**, 175-190.
- [6] Barrodale I. and Roberts F. D. K. (1973). An improved algorithm for



- discrete L1 linear approximation. *SIAM J. Numer. Anal.* **10**, 839-848.
- [7] Barrodale I. and Roberts F. D. K. (1974). Algorithm 478. Solution of an overdetermined system of equations in the L1 norm. *Commun. ACM* **14**, 319-320.
  - [8] Beasley (1990). Linear programming on Cray supercomputers. *J. Opl. Res. Soc.* **41**, 133-139.
  - [9] Beaton A. E. and Tukey J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band spectrographic data. *Technometrics* **16**, 147-185.
  - [10] Bloomfield P. and Steiger W. L. (1980). Least absolute deviations curve fitting. *SIAM J. Scient. Statist. Comput.* **1**, 290-301.
  - [11] Bloomfield P. and Steiger W. L. (1984). *Least Absolute Deviations: Theory, Applications and Algorithms*. Boston: Birkhauser.
  - [12] Charnes A. Cooper W. W. and Ferguson R. O. (1955). Optimal estimation of executive compensation by linear programming. *Management Science* **1**, 138-151.
  - [13] Dielman T. E. (1992). Computational algorithms for least absolute value regression. In *L1 Statistical Analysis and Related Methods*, Ed. Y. Dodge, pp. 311-326. Elsevier Science Publishers.
  - [14] Dodge Y., Antoch J. and Jureckova J. (1991). Adaptive combinations of least squares and least absolute deviations estimators: computational aspects. *Comput. Statist Data Anal.* **12**, 87-99.
  - [15] Dodge Y. and Jureckova J. (1991). Flexible L-estimation in the linear Model. *Comput. Statist Data Anal.* **12**, 211-220.
  - [16] Dodge Y. and Jureckova J. (1992). A Class of estimators based on adaptive convex combinations of two estimation procedures. In *L1 Statistical Analysis and Related Methods*, Ed. Y. Dodge, pp. 31-45. Elsevier Science Publishers.
  - [17] Fair R. C. (1974). On the robust estimation of econometric models. *Ann. Econ. Soc. Measurement* **3**, 667-677.
  - [18] Schlossmacher E. J. (1973). An iterative technique for absolute deviations curve fitting. *J Am. Statist. Soc* **68**, 857-865.
  - [19] Seneta E. and Steiger W. L. (1984). A new LAD curve fitting algorithm: slightly over-determined equation systems in L1. *Discrete Applied Mathematics* **7**, 79-91.

# Exact algorithms for computing the least median of squares estimate in multiple linear regression

José Agulló

*Universidad de Alicante, Spain*

*Abstract:* We propose two finite algorithms to compute the exact least median of squares (LMS) estimates of parameters of a linear regression model with  $p$  coefficients. The first algorithm is similar to Stromberg's (1993) exact algorithm. It is based on the exact fit to subsets of  $p$  cases and uses impossibility conditions to avoid unnecessary calculations. The second one is based on a branch and bound (BAB) technique. Empirical results suggest that the proposed algorithms are faster than the finite exact algorithms described earlier in the literature.

*Key words:* Branch and bound, exact algorithms, high breakdown regression, least median of squares, robust regression

AMS subject classification: 62F35, 62J05, 65U05.

## 1 Introduction

In this paper we consider the multiple linear regression model

$$Y = Z\theta + \epsilon, \tag{1}$$

where  $Y$  is an  $n \times 1$  vector of dependent variables,  $\theta$  is a  $p \times 1$  vector of unknown parameters,  $Z$  is an  $n \times p$  design matrix of predictors, and  $\epsilon$  is an  $n \times 1$  vector of true residuals. We denote the  $i$ th component of  $Y$  and the  $i$ th row of  $Z$  by  $y_i$  and  $z_i^t$ , respectively. We suppose that the design matrix  $Z$  is fixed and has rank  $p$ . Sometimes we also assume that any  $p \times p$  submatrix of  $Z$  is nonsingular; in this case we say that  $Z$  verifies the *Haar condition*, or that the observations are in *general position*. An estimate of  $\theta$ , say  $\hat{\theta}$ , gives  $n$  residuals  $e_i(\hat{\theta}) = y_i - z_i^t \hat{\theta}$ . The most well-known

estimator of  $\theta$  is the Least Squares (LS) estimator. The LS estimator is optimal in several situations but it is severely affected by outliers. It also suffers from the *problem of masking*, which occurs when a data set contains multiple outliers and, at the same time, these outliers are not detected by the usual LS diagnostic procedures. To get a reliable outlier detection and estimation, a high breakdown point estimator should be used. Such an estimator is the *Least Median of Squares* (LMS) estimator, introduced by Rousseeuw (1984). The LMS estimator, which we denote by  $\hat{\theta}_{LMS}$ , is defined by

$$\hat{\theta}_{LMS} = \operatorname{argmin}_{\theta} \{|e_i(\theta)|\}_{h:n},$$

where the subscript  $h:n$  denotes the  $h$ th order statistic of  $n$  numbers. To achieve the maximum breakdown point when the observations are in general position, the coverage  $h$  is chosen as  $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ , where  $\lfloor \cdot \rfloor$  denotes the greatest integer function. For  $h = n$ , the LMS estimator is equal to the *Chebyshev estimator*, denoted by  $\hat{\theta}_C$ , and defined by

$$\hat{\theta}_C = \operatorname{argmin}_{\theta} \{\max_i |e_i(\theta)|\}.$$

The minimal value of the objective function which defines  $\hat{\theta}_C$  will be called the *Chebyshev criterion*.

The computation of the LMS estimate is difficult. The most widely used algorithm to approximate the LMS estimate is the PROGRESS algorithm (Rousseeuw and Leroy, 1987). This algorithm computes the exact fit of several *elemental sets* (subsets of size  $p$ ) of the data set. The exact fit with smallest  $h$ th absolute residual gives an approximate LMS estimate. The considered elemental sets can be either all  $\binom{n}{p}$  possible such sets, or a random subsample of them. If the regression model has intercept, the intercept of each exact fit can be adjusted to yield a smaller  $h$ th absolute residual. For a simple regression model ( $p = 2$ ), Steele and Steiger (1986) prove that an algorithm which examines all elemental sets and adjusts the intercept for each of them, obtains the exact LMS estimate. However, for  $p$  greater than 2 such an algorithm does not necessarily yield the exact LMS estimate.

Stromberg (1993) proposes an exact algorithm based on the fact that the exact LMS estimate is a Chebyshev estimate for some subset of size  $p + 1$  of the data. Any set of  $p + 1$  indices from  $\{1, \dots, n\}$  will be called here a *reference set*. Stromberg's algorithm examines all reference sets, computes the Chebyshev estimate for each of them, and then sets  $\hat{\theta}_{LMS}$  as the Chebyshev estimate with smallest  $h$ th absolute residual.

In this paper we develop two algorithms to compute the exact LMS estimate that are computationally feasible for small or moderate size samples.

Section 2 describes some properties of the Chebyshev fit and discusses its computation. In Section 3 we propose an algorithm to find the optimal reference set. This algorithm differs from Stromberg's in some important features: It is based on exact fits of elemental sets; it uses accelerations based on impossibility conditions, and it also avoids completely sorting of residuals. In Section 4 we propose a finite exact algorithm based on a *branch and bound* (BAB) technique. The BAB algorithm finds the subset of size  $h$  whose Chebyshev estimate gives the exact LMS estimate without exhaustive enumeration of all  $h$ -subsets. We describe the basic BAB algorithm and some strategies to improve its efficiency. An empirical comparison of the LMS algorithms is described in Section 5.

## 2 Chebyshev regression

In this Section we describe some properties of the Chebyshev fit and discuss its computation. Osborne and Watson (1967) proved that, if the design matrix  $Z$  of model (1) has rank  $p$  and  $n > p$ , then: First, there exists a Chebyshev estimate that is equal to a Chebyshev estimate for some reference set of the data; second, the Chebyshev criterion is identical to the Chebyshev criterion for the optimal reference set; third, all cases of the optimal reference set have residuals whose absolute value is equal to the Chebyshev criterion, and fourth, the rank of the design matrix of the optimal reference set is  $p$ . It is well known that the Haar condition is sufficient (but not necessary) for uniqueness of the Chebyshev estimate.

For a reference set, both the Chebyshev criterion and a Chebyshev estimate can be explicitly computed. Assume that  $n = p + 1$  and the design matrix  $Z$  verifies the Haar condition. In this case two methods to obtain the Chebyshev estimate are available. The first method (see, e.g., Cheney, 1966; Meicler, 1968) is based on the LS estimate  $\hat{\theta}_{LS} = (Z^t Z)^{-1} Z^t Y$ . Let  $\hat{e}^t = (\hat{e}_1, \dots, \hat{e}_{p+1})$  be the LS residual vector, and denote  $s^t = (\text{sgn}(\hat{e}_1), \dots, \text{sgn}(\hat{e}_{p+1}))$ , where "sgn" represents the sign function. Then the Chebyshev criterion is

$$\omega = \begin{cases} 0 & \text{if } \sum_{i=1}^{p+1} \hat{e}_i^2 = 0 \\ \sum_{i=1}^{p+1} \hat{e}_i^2 / \sum_{i=1}^{p+1} |\hat{e}_i| & \text{otherwise} \end{cases} \quad (2)$$

and the Chebyshev estimate is

$$\hat{\theta}_C = \hat{\theta}_{LS} - \omega(Z^t Z)^{-1} Z^t s = (Z^t Z)^{-1} Z^t (Y - \omega s). \quad (3)$$

The second method (see, e.g., Meicler, 1969; Armstrong and Kung, 1980) is based on the exact fit to one of  $p + 1$  possible elemental sets. The set

$\{1, \dots, p+1\}$  is partitioned into an elemental set  $J$  and its complement. Suppose (without loss of generality) that  $J$  contains the first  $p$  indices, and let  $r = p+1$  be the complementary index. Denote by  $Z_J$  and  $Y_J$  the respective submatrices of  $Z$  and  $Y$  formed with the rows indexed by  $J$ . As  $Z_J$  is nonsingular,  $\hat{\theta}_J = Z_J^{-1}Y_J$  is the exact fit to the elemental set  $J$ , and  $e_i = y_i - z_i^t \hat{\theta}_J$  is the  $i$ th residual based on the exact fit. Let  $\xi = (\xi_1, \dots, \xi_p) = z_r^t Z_J^{-1}$ , and  $\sigma^t = -\text{sgn}(e_r)(\text{sgn}(\xi_1), \dots, \text{sgn}(\xi_p))$ . Then the Chebyshev criterion is

$$\omega = \frac{|e_r|}{1 + \sum_{j=1}^p |\xi_j|}, \quad (4)$$

and the Chebyshev estimate is

$$\hat{\theta}_C = \hat{\theta}_J - \omega Z_J^{-1} \sigma = Z_J^{-1}(Y_J - \omega \sigma). \quad (5)$$

**Remark 1** *Although for a given reference set both methods give the same  $\omega$  and  $\hat{\theta}_C$ , an implementation based on the second method is more convenient to compute the exact LMS estimate. To our knowledge, no exact algorithm to compute the LMS estimator using the second method has yet appeared in the literature. The exact LMS algorithm of Stromberg (1993) computes the Chebyshev estimates of reference sets using the first method.*

**Remark 2** *Assume  $n = p+1$  and  $\omega > 0$ . If the matrix  $Z$  has rank  $p$  but it does not verify the Haar condition, then there exist multiple Chebyshev estimates (see, e.g., Cheney, 1966, p. 42, problems 6, 7). For our purposes, the search can be restricted to those Chebyshev estimates for which all  $p+1$  residuals are equal in magnitude to  $\omega$ . Using the first method, if multiple Chebyshev estimates exist, at least one  $\hat{e}_i$  will be equal to zero. For these null residuals, the respective components of  $s$  may be set equal to 1 or -1, and using (3), multiple Chebyshev estimates are obtained. In the second method, an elemental set  $J$  with nonsingular design matrix has to be selected from the reference set. For this method, when multiple Chebyshev estimates exist, some  $\xi_i$  will be equal to zero. Changing conveniently the values assigned to the signs of such  $\xi_i$ 's, expression (5) yields multiple Chebyshev estimates.*

When  $n > p+1$ , a naive algorithm computing the Chebyshev estimate examines all possible reference sets: For each reference set, it computes the Chebyshev criterion  $\omega$  through (2) or (4), and selects the reference set which yields the greatest  $\omega$ . Afterwards, it computes the Chebyshev estimate using (3) or (5). However, several nonexhaustive algorithms have been proposed in the literature for this purpose. These algorithms are based

on a linear programming formulation of the Chebyshev regression problem that consists of

$$\text{Minimize } \omega, \quad \text{s. t.: } -\omega \leq y_i - z_i^t \theta \leq \omega, \quad i = 1, \dots, n. \quad (6)$$

We prefer the algorithm of Armstrong and Kung (1980) (AK) which uses the dual form of (6). The source code of a FORTRAN implementation of the AK algorithm appears in Armstrong and Kung (1979). (The published code contains one harmful misprint which is corrected in Agulló, 1994). The AK algorithm obtains a finite sequence of reference sets that converges to the optimal reference set. In this sequence, consecutive reference sets only differ in one index. In each iteration, the index to be added to the current reference set would be that associated with the maximum absolute residual. The index to be dropped is selected using a rule that increases the objective function value. Notice that the AK algorithm obtains in each iteration a lower and an upper bound for the Chebyshev criterion. The lower and upper bounds are, respectively, the objective function value and the maximum absolute residual. When both bounds are identical, convergence is achieved.

### 3 Exhaustive LMS algorithm

Assume that matrix  $Z$  in model (1) verifies the Haar condition. Since the exact LMS estimate minimizes the  $h$ th smallest absolute residual, it must minimize the maximum absolute residual for some subset of size  $h$  of the data (i.e.,  $\hat{\theta}_{LMS}$  is the Chebyshev estimate to some  $h$ -subset). Moreover, from the properties of the Chebyshev regression (see Section 2), we conclude that the exact LMS estimate is equal to the Chebyshev estimate of some reference set. Further, the minimized  $h$ th absolute residual is the same as the Chebyshev criterion of this optimal reference set. Consequently, in order to compute the exact LMS estimate we can use an exhaustive algorithm that examines all reference sets. For each reference set the algorithm computes the Chebyshev estimate, and sets  $\hat{\theta}_{LMS}$  as the Chebyshev estimate with smallest  $h$ th absolute residual.

It is possible to carry out some fast preliminary tests to discard those reference sets which cannot be optimal. The tests are based on the following fact. Suppose we know that the optimal  $h$ th absolute residual is not greater than  $\omega^*$ . Then a reference set (with Chebyshev criterion  $\omega$  and Chebyshev estimate  $\hat{\theta}_C$ ) cannot be optimal if it verifies any of the following three

*impossibility conditions:*

$$\begin{aligned}
 [I] \quad & \omega > \omega^*, \\
 [II] \quad & \#\{1 \leq i \leq n, |e_i(\hat{\theta}_C)| > \omega\} > n - h, \\
 [III] \quad & \#\{1 \leq i \leq n, |e_i(\hat{\theta}_C)| > \omega^*\} > n - h,
 \end{aligned}$$

where we use the symbol  $\#$  to denote the cardinal of a set.

Stromberg's exact algorithm examines all reference sets, and computes the Chebyshev estimate  $\hat{\theta}_C$  for each one using (3). It uses impossibility condition [III] to avoid the computation of all absolute residuals and/or their sorting for some Chebyshev estimates. When a reference set does not verify the impossibility condition [III], its Chebyshev estimate becomes the potentially best estimate and the absolute residuals are sorted to find the  $h$ th smallest absolute residual. This absolute residual becomes  $\omega^*$ .

We describe next an exact LMS algorithm based also on an exhaustive enumeration of reference sets. It uses impossibility conditions [I] and [II] to avoid unnecessary calculations. Initially the algorithm sets  $\omega^* := \infty$ , and  $R^* := \emptyset$ . Then it considers all elemental sets  $J = \{j_1, \dots, j_p\} \subset \{1, \dots, n\}$  with  $1 \leq j_1 < \dots < j_p \leq n - 1$ . For each elemental set  $J$ , it computes the exact fit  $\hat{\theta}_J = Z_J^{-1}Y_J$ . Then it examines the reference sets formed by adding an index  $r$  to  $J$ , where  $j_p < r \leq n$ . For each reference set  $R = J \cup \{r\}$ , the Chebyshev criterion  $\omega$  is computed using (4). If  $R$  verifies impossibility condition [I], then  $R$  cannot be optimal and it is discarded. Otherwise, the algorithm computes the Chebyshev estimate  $\hat{\theta}_C$  of  $R$  using (5), and calculates the Chebyshev residuals until either a) the number of absolute residuals greater than  $\omega$  equals  $n - h + 1$ , or b) the number of absolute residuals that are not greater than  $\omega$  becomes equal to  $h$ . When a) occurs,  $R$  cannot be optimal (because  $R$  verifies impossibility condition [II]) and it is discarded. When b) occurs, the algorithm sets  $\omega^* := \omega$ ,  $R^* := R$ , and  $\hat{\theta}^* := \hat{\theta}_C$ . When the algorithm stops,  $\omega^*$  yields the optimal criterion,  $R^*$  the optimal reference set, and  $\hat{\theta}^*$  the exact LMS estimate.

Matrix inversion is not needed to implement this algorithm using, for instance, an LU decomposition of  $Z_J$ . Notice that the same LU decomposition is used for evaluating, on average,  $n/(p + 1)$  reference sets and it is only updated when the elemental set changes.

Our algorithm differs from Stromberg's in many important respects. In Stromberg's algorithm the Chebyshev fits are based on the LS fits of reference sets. However, in our proposal the Chebyshev fits are based on exact fits of elemental sets. So, our algorithm uses the same exact fit to examine several Chebyshev fits, and it can be implemented adapting the available LMS algorithms based on elemental sets. A further benefit of our approach

is that the computation of an approximate high breakdown multivariate estimate for matrix  $Z$  (such as Rousseeuw's (1985) minimum volume ellipsoid estimate) can be done at the same time as the exact LMS computation with little additional cost (see Hawkins and Simonoff, 1993). Stromberg's algorithm requires the computation of the Chebyshev estimators for all reference sets. However, our algorithm uses impossibility condition  $[I]$  to avoid the computation of the Chebyshev estimate and the Chebyshev residuals for a significant fraction of all reference sets. Moreover, when the computation of Chebyshev residuals has to be started, it seldom requires to compute all residuals and always completely avoids sorting of residuals.

The proposed algorithm examines  $\binom{n}{p+1} = O(n^{p+1})$  reference sets. At its worst, it must compute  $n$  residuals for each reference set, therefore requiring at most  $O(n^{p+2})$  time. On the other hand, Stromberg's algorithm requires  $O(n^{p+2} \log n)$  time, but if the  $h$ th smallest absolute residual is found in  $O(n)$  time, it also requires  $O(n^{p+2})$  time. Notice that although both algorithms have the same computational complexity, this does not imply that their true computer times are the same, since the involved constant factors can be very different. Empirical results suggest that our proposal is about five times faster than Stromberg's algorithm (see Section 5).

**Remark 3** *To analyse data sets whose design matrix has rank  $p$  and does not necessarily verify the Haar condition, the proposed algorithm requires two modifications. The first modification is concerned with the possibility of any nonsingular matrix  $Z_J$ . For each elemental set  $J$ , the numerical rank of  $Z_J$  should be checked. If the rank of  $Z_J$  is  $p$  no problem arises. When the rank of  $Z_J$  is smaller than  $p - 1$ , the next elemental set in the list is examined. If both matrices  $Z_J$  and  $Z_R$  have rank  $p - 1$ , the next reference set in the list is examined. Finally, if the rank of  $Z_J$  equals  $p - 1$  and the rank of  $Z_R$  is  $p$ , then the algorithm selects an elemental set from  $R$  whose design matrix has rank equal to  $p$ . Now, the complementary index of the elemental set plays the role of  $r$ . The second modification deals with the possibility of having some  $R$  with multiple Chebyshev estimates. The multiplicity of Chebyshev estimates occurs when some  $\xi_i$  in (4) is equal to zero. If this occurs, we must examine all Chebyshev estimates whose residuals indexed by  $R$  are equal in magnitude to  $\omega$  (see Remark 2).*

## 4 Branch and bound algorithm

As we noted earlier, the exact LMS estimate coincides with the Chebyshev estimate of some  $h$ -subset. In fact, the optimal  $h$ -subset is the one with smallest Chebyshev criterion. In this Section we propose a Branch And



Bound (BAB) algorithm to find this optimal  $h$ -subset.

Given an index set  $J_m = (j_1, \dots, j_m) \subset \{1, \dots, n\}$  with size  $m$ , let  $Z_{J_m}$  and  $Y_{J_m}$  be, respectively, the submatrices of  $Z$  and  $Y$  formed by the rows indexed by  $J_m$ . For the regression of  $Y_{J_m}$  on  $Z_{J_m}$ , we denote the sum of squared LS-residuals by  $\phi(J_m)$ , the sum of absolute values of LS-residuals by  $\phi'(J_m)$ , and the Chebyshev criterion by  $\omega(J_m)$ . It is easy to prove that:

$$J \subset J' \implies \omega(J) \leq \omega(J'), \quad (7)$$

$$B(J_m) \leq \omega(J_m), \quad (8)$$

$$\phi(J_m) > 0 \implies B(J_m) \leq B'(J_m) \leq \omega(J_m), \quad (9)$$

where  $B(J_m) = \sqrt{\phi(J_m)/m}$  and  $B'(J_m) = \phi(J_m)/\phi'(J_m)$ .

Note that the monotonicity property (7) implies that the Chebyshev criterion cannot decrease if one or several cases are added to the current set of cases.

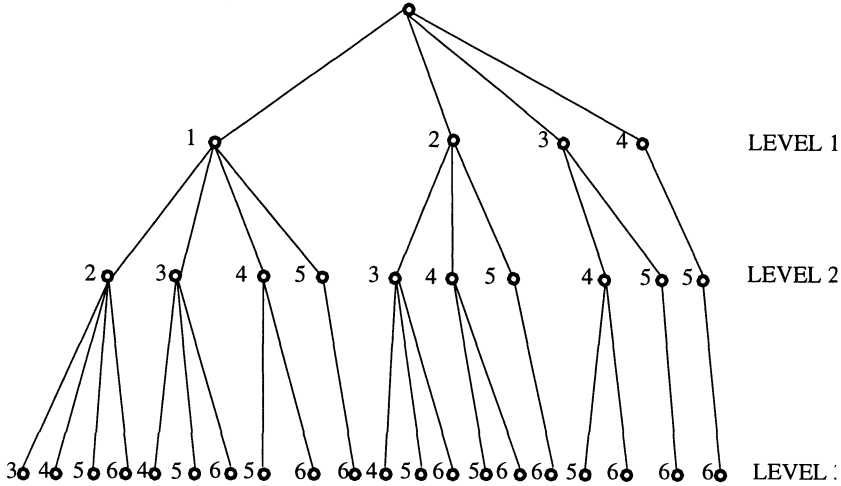
In the search of the optimal  $h$ -subset, the BAB algorithm considers subsets whose size is not greater than  $h$ . The current subset will be denoted by  $J_m = (j_1, \dots, j_m)$ . We start by considering sequences that verify

$$j_1 < \dots < j_m. \quad (10)$$

The generation of subsets is organized through a tree of nested subsets with  $h$  node levels. In each node, a further index is added from the original  $n$  indices. We use the tree described in Narendra and Fukunaga (1977). Figure 1 shows a tree for  $n = 6$  and  $h = 3$ . A node at level  $m$  is labeled with the value of  $j_m$  and represents a subset of size  $m$ . Terminal nodes represent the  $\binom{n}{h}$  possible subsets of  $h$  observations. When an exhaustive inspection of the tree is carried out, the tree is examined by moving down each branch, working from right to left. When a terminal node is reached, the inspection continues from the most recent node that has unexplored branches.

The efficiency of the BAB algorithm follows from the possibility of jumping in the exhaustive inspection sequence of the tree. In any stage of the search, let  $\omega^*$  be the smallest Chebyshev criterion for a  $h$ -subset so far obtained. If the subset being considered is  $J_m$ ,  $1 \leq m \leq h$ , and if it verifies that  $\omega(J_m)$  is greater than  $\omega^*$ , then, as a consequence of the monotonicity property, all  $h$ -subsets that contain  $J_m$  can be rejected implicitly.

Figure 1: Tree for  $n = 6$  and  $h = 3$ . Labels at nodes denote the case that is added there.



We describe now the basic BAB algorithm. Initially it sets  $\omega^* := \infty$  and  $J^* := \emptyset$ . If the current subset  $J_m$  verifies that  $B(J_m)$  is greater than  $\omega^*$ , then, from (7) and (8), the optimal  $h$ -subset cannot contain  $J_m$ , and the exploration can be continued from the most recent node that has unexplored branches. In this case, a jump occurs in the exhaustive sequence. When the current node is terminal (i.e.,  $m = h$ ) and it verifies that  $B(J_h)$  is not greater than  $\omega^*$ , the bound  $B'(J_h)$  is computed. If  $B'(J_h) \geq \omega^*$ , by (7) and (9),  $J_h$  cannot be optimal, and the exploration of the tree continues. When  $B'(J_h) < \omega^*$ , the iterative computation of  $\omega(J_h)$  (using the AK algorithm) starts. If, in some iteration of this computation, the objective function value is greater than or equal to  $\omega^*$  (what implies  $\omega(J_h) \geq \omega^*$ ), then the iterative process stops and the exploration of the tree continues. When the AK algorithm converges (i.e.,  $\omega(J_h)$  is smaller than  $\omega^*$ ), the BAB algorithm sets  $\omega^* := \omega(J_h)$  and  $J^* := J_h$ , and the exploration of the tree continues. When the BAB algorithm stops,  $J^*$  is the optimal  $h$ -subset and  $\omega^*$  is the minimal value of the LMS objective function.

The computation of the bound  $B$  requires the sum of squared LS-residuals. This computation is carried out by using an orthogonal decomposition procedure (Gentleman, 1974) applied to  $(Z_J, Y_J)$ . When a case is added to the current subset, the orthogonal factors are updated. When the algorithm operates descending by a branch of the tree, the orthogonal factors of each level are saved. In this way, the algorithm can reselect the adequate factors when it returns to a smaller level node. This permits updating the orthogonal factors when the inspection continues by an unex-

plored branch. When a terminal node is reached and it verifies  $B(J_h) < \omega^*$ , then the computation of the bound  $B'(J_h)$  is required. This computation is carried out quickly from the orthogonal factors. If  $B'(J_h) > \omega^*$ , the computation of  $\omega(J_h)$  is avoided, because  $J_h$  is not optimal.

We describe now two strategies to improve the computational efficiency of the basic BAB algorithm. Suppose that  $m < h$ ,  $\text{rank}(Z_{J_m}) = p$ , and  $0 < B(J_m) < \omega^*$ . Since the orthogonal decomposition of  $(Z_J, Y_J)$  is available, the bound  $B'(J_m)$  can be quickly computed. If  $B'(J_m) \geq \omega^*$ , then, by (7) and (9), a jump is justified. Notice that if  $m = p + 1$ , then  $\omega(J_m) = B'(J_m)$ , by (2), whereas if  $m > p + 1$ , then  $\omega(J_m) \geq B'(J_m)$ . If  $m > p + 1$  and  $B'(J_m) < \omega(J_m)$ , then the iterative AK algorithm to compute  $\omega(J_m)$  can be started and continued until either a) the current objective function value is greater than or equal to  $\omega^*$ , or b) the current greatest absolute residual is smaller than  $\omega^*$ . From the properties of the AK algorithm (see Section 2), when a) occurs, we conclude that  $\omega(J_m)$  is greater than  $\omega^*$ , and consequently, a jump in the exhaustive sequence is justified. However, when b) occurs,  $\omega(J_m)$  is smaller than  $\omega^*$ , and a jump cannot be justified.

In a node  $j_m$  at level  $m$  with ancestor nodes  $j_1, \dots, j_{m-1}$  only certain indices can be selected as successors in level  $m + 1$ . All indices selected at nodes with labels  $j_1, \dots, j_m$  and all indices that appear in the “brother” nodes to the left of nodes  $j_1, \dots, j_m$  are not available. Assume that the current node has  $n_s$  successors at level  $m + 1$ . The basic BAB algorithm selects the first  $n_s$  available indices and assigns them to the successors by order from left to right in the tree. If we do not impose restriction (10) in the subset generation process, we can use a *sorting rule* for assigning available indices to the successors. Suppose that the current design matrix  $Z_{J_m}$  has rank  $p$ . For each available index  $i$  we can compute the increase in the sum of squared LS-residuals caused by adding the index  $i$  to  $J_m$ . This increase, which we call *ith partial increment*, is

$$\gamma_i = \frac{y_i - z_i^t (Z_{J_m}^t Z_{J_m})^{-1} Z_{J_m}^t Y_{J_m}}{1 + z_i^t (Z_{J_m}^t Z_{J_m})^{-1} z_i}.$$

The sorting rule selects the  $n_s$  available indices with greatest partial increments and assigns them to the successors according to the decreasing magnitude of the partial increment from left to right in the tree. If the BAB algorithm uses that sorting rule and the current subset  $J_m$  verifies that  $B(J_m)$  is greater than  $\omega^*$ , then the subtree emanated from the current node does not have to be examined. This follows from (7) and (8). Furthermore, if the design matrix of the previous level has rank  $p$ , as a consequence of the sorting rule all brother nodes to the left of the current node and the subtrees descending from such nodes can also be discarded.

**Remark 4** Denote  $\gamma^* = (m+1)(\omega^*)^2 - \phi(J_m)$ , and let  $\gamma' = \gamma_{n_s, n_a}$  be the  $n_s$ th greatest partial increment from the  $n_a$  available indices. If  $\gamma'$  is greater than  $\gamma^*$ , it is not necessary to assign indices to successors, because the optimal  $h$ -subset cannot contain the current subset  $J_m$ . Sometimes the computation of partial increments for all available indices and its partial sorting can be avoided. Suppose that we proceed to compute the partial increments and maintain simultaneously a counter,  $n_r$ , to determine the number of partial increments that are not smaller than  $\gamma^*$ . If, at some stage of this process,  $n_r$  equals  $n_s$ , then  $\gamma'$  will necessarily be greater than  $\gamma^*$ , and a jump in the exhaustive sequence is justified.

Even if the tree is enumerated through the sorting rule, the efficiency of the BAB algorithm depends on the initial ranking of cases in the data set. Notice that for nodes at low levels, the sorting rule assigns the first available indices to the successors. The empirical tests we have carried out suggest the convenience of reassigning the indices  $1, \dots, n$  to the cases according to the decreasing magnitude of residuals based on an approximate LMS estimate. In this way, the first  $h$ -subset evaluated by the BAB algorithm gives the approximate LMS estimate and a good upper bound for the optimal value of the objective function is found quickly. The approximate LMS estimate is computed by the BAB algorithm before starting to inspect the tree. For this task, we use a modification of the Feasible Subset Algorithm (FSA) described in Hawkins (1993).

## 5 Empirical comparison of the LMS algorithms

To compare the efficiency of the LMS algorithms, we have considered the following FORTRAN implementations:

- MVELMS, code of Hawkins and Simonoff (1993). We use the option that examines all elemental sets and adjusts the intercept for each elemental set.
- EXTLMS, implementation of Stromberg's exact algorithm due to Hawkins, Simonoff and Stromberg (1994).
- LULMS, our implementation of the algorithm described in Section 3. It uses an LU decomposition of the basis as described by Bartels and Golub (1969) to obtain solutions of square systems. The used code implements the first modification described in Remark 3, but not the second one.
- MVELMS1, our modification of the MVELMS code. This modification is based on the algorithm explained in Section 3, and omits from the search those elemental sets whose design matrix is singular.
- BABLMS, our implementation of the BAB algorithm described in Sec-

tion 4. It obtains the exact LMS estimate without exhaustive enumeration of  $h$ -subsets.

Table 1: Comparison of CPU times (in seconds) for LMS algorithms.

| DATA SET | $n$ | $p$ | $h$ | CPU TIME (in seconds) |         |         |        |        |
|----------|-----|-----|-----|-----------------------|---------|---------|--------|--------|
|          |     |     |     | EXTLMS                | MVELMS1 | LULMS   | MVELMS | BABLMS |
| Aircraft | 23  | 5   | 14  | 42.69                 | 13.30   | 8.30    | 8.35   | 0.66   |
| Coleman  | 20  | 6   | 13  | 43.13                 | 9.95    | 7.09    | 10.44  | 0.32   |
| Delivery | 25  | 3   | 14  | 2.97                  | 0.88    | 0.50    | 0.44   | 0.11   |
| Educat   | 50  | 4   | 27  | 804.18                | 343.08  | 159.01  | 102.31 | 10.82  |
| Hawkins  | 75  | 4   | 39  | 7774.73               | 4823.85 | 1905.11 | 834.07 | 531.31 |
| Races    | 35  | 3   | 19  | 13.46                 | 4.01    | 2.25    | 1.75   | 0.27   |
| Salinity | 28  | 4   | 16  | 31.92                 | 8.85    | 5.38    | 5.16   | 0.43   |
| Stacklos | 21  | 4   | 12  | 6.48                  | 1.65    | 1.21    | 1.15   | 0.11   |
| Wood     | 20  | 6   | 13  | 43.02                 | 9.78    | 7.09    | 10.33  | 0.38   |
| Total    |     |     |     | 8762.58               | 5215.35 | 2095.94 | 974.00 | 544.90 |

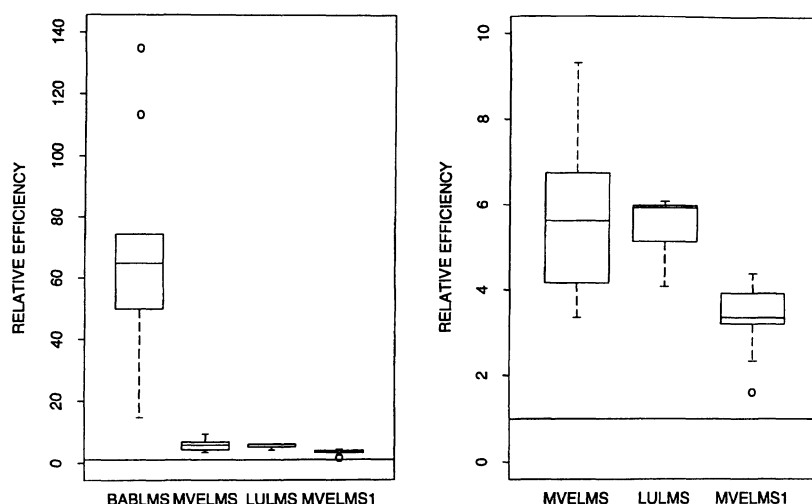
From these algorithms, MVELMS gives only an approximate LMS estimate, whereas the remaining algorithms are exact. EXTLMS, LULMS and MVELMS1 guarantee an exact LMS estimate when the observations are in general position, and BABLMS obtains a true LMS estimate when the design matrix has rank  $p$ , which is a much weaker condition. In our implementations, we use subroutines of Armstrong and Kung (1979), Miller (1992), Miller and Nguyen (1994), Ridout (1988), and Wichmann and Hill (1982). All source codes were compiled with FTN77 for 486 and were run on a 90 MHz Pentium computer. The central processor unit (CPU) times for several data sets are shown in Table 1. All data sets, excluding the one labeled “Races”, can be found in Rousseeuw and Leroy (1978). The Races data set is from Atkinson (1986). Figure 2 shows the multiple boxplot of relative efficiencies, measured using the ratio of CPU time to EXTLMS CPU time.

Table 1 and Figure 2 show that for the used data sets the BABLMS algorithm dominates all other algorithms. On average, the BABLMS algorithm is about 65 times faster than the EXTLMS one, and at least one order of magnitude faster than the exhaustive MVELMS algorithm that is approximate. This may be surprising, because BABLMS searches a subset within a collection of size  $\binom{n}{h}$ , and this number will, in general, be much greater than, for instance,  $\binom{n}{p+1}$ , which is the number of subsets inspected by EXTLMS. The reason why the BABLMS algorithm is more efficient is because the number of subsets explicitly visited by it is usually smaller than the number of reference sets.

Notice that the exact LULMS algorithm is about five times faster than

the EXTLMS algorithm, and it does not need much more computer time than the approximate MVELMS algorithm based on exhaustive search over all elemental sets.

Figure 2: Multiple boxplots of relative efficiency of LMS algorithms. The baseline corresponds to the EXTLMS algorithm.



To sum up, we can conclude that for small or moderate sample sizes the proposed algorithms are more efficient than the other finite exact algorithms proposed in the literature.

## References

- [1] Agulló, J. (1994). A remark on algorithm AS 135. Mimeo.
- [2] Armstrong, R.D., and Kung, D.S. (1979). Algorithm AS135. Min-max estimates for a linear multiple regression problem. *Appl. Statist.* **28** 93–100.
- [3] Armstrong, R.D., and Kung, D. (1980). A dual method for discrete Chebychev curve fitting. *Mathematical Programming* **19** 186–199.
- [4] Atkinson, A.C. (1986). Aspects of diagnostic regression analysis. *Statistical Science* **1** 397–402.
- [5] Bartels, R.G., and Golub, G.H., (1969). The simplex method of linear programming using LU decomposition, *Commun. of the ACM* **12** 266–268.
- [6] Cheney, E.W. (1966). *Introduction to Approximation Theory*. New York: McGraw-Hill.
- [7] Gentleman, W.M. (1974). Algorithm AS75. Basic procedures for large, sparse or weighted linear least squares problems. *Appl. Statist.* **23**

448–454.

- [8] Hawkins, D.M. (1993). The feasible set algorithm for least median of squares regression. *Comput. Statist. and Data Anal.* **16** 81–101.
- [9] Hawkins, D.M., and Simonoff, J.S. (1993). Algorithm AS 282. High breakdown regression and multivariate estimation. *Appl. Statist.* **42** 423–441.
- [10] Hawkins, D.M., Simonoff, J.S., and Stromberg, A. (1994). Distributing a computationally intensive estimator: the case of exact LMS regression. *Comput. Statist.* **9** 83–95.
- [11] Meicler, M. (1968). Chebyshev solution of an inconsistent system of  $n+1$  linear equations in  $n$  unknowns in terms of its least-squares solution. *SIAM Rev.* **10** 373–375.
- [12] Meicler, M. (1969). A steepest ascent method for the Chebyshev problem. *Mathematics of Computation* **23** 813–817.
- [13] Miller, A.J. (1992). Algorithm AS 274. *Appl. Statist.* **41** 458–478.
- [14] Miller, A.J., and Nguyen, N. (1994). Algorithm AS 295. A Fedorov exchange algorithm for D-optimal design. *Appl. Statist.* **43** 669–678.
- [15] Narendra, P.M., and Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers* **26** 917–922.
- [16] Osborne, M.R., and Watson, G.A. (1967). On the best linear Chebyshev approximation. *Computer Journal* **10** 172–177.
- [17] Ridout, M.S. (1988). Algorithm AS 233. An improved branch and bound algorithm for feature subset selection. *Appl. Statist.* **37** 139–147.
- [18] Rousseeuw, P.J. (1984). Least median of squares regression. *J. Am. Statist. Soc.* **79** 871–880.
- [19] Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications, B*, Eds. W. Grossmann, G. Pflug, and W. Wertz, pp. 283–297. Dordrecht: Reidel Publishing.
- [20] Rousseeuw, P.J., and Leroy, A. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
- [21] Steele, J.M., and Steiger, W.L. (1986). Algorithms and complexity for LMS regression. *Discrete Applied Mathematics* **14** 93–100.
- [22] Stromberg, A. (1993). Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM J. Scient. Comput.* **14** 1289–1299.
- [23] Wichman, B.A., and Hill, I.D. (1982). Algorithm AS 183. An efficient and portable pseudorandom number generator. *Appl. Statist.* **31** 188–190.

# Polynomial algorithms for isotonic regression

**Victor Chepoi**

*Universitatea de stat din Moldova, Chişinău, Moldavia and  
Université d'Aix Marseille II, France*

**Daniel Cogneau and Bernard Fichet**

*Université d'Aix Marseille II, France*

*Abstract:* In this paper we consider the following problem. Let  $X = \{x_1, \dots, x_n\}$  be a set of observation points endowed with a partial order  $\prec$ , and let  $y_1, \dots, y_n$  be the values of the dependent variable  $y$ . We are searching an isotonic function  $f : X \rightarrow \mathbb{R}$  (i.e.  $x_i \prec x_j$  implies that  $f(x_i) \leq f(x_j)$ ) that minimizes the  $l_p$ -error

$$D_p(f) = \left[ \sum_{x_i \in X} |y_i - f(x_i)|^p \right]^{\frac{1}{p}}.$$

We recall some general algorithms for solving this and related regression problems and we present new polynomial algorithms for some versions of the isotonic regression problem.

*Key words:*  $L_p$ -norm, polynomial algorithm, curve fitting, graphs, partial ordering.

AMS subject classification: 62J05.

## 1 Introduction

The basic isotonic regression problem can be formulated as follows: given values  $y_1, \dots, y_n$  of the dependent variable  $y$ , corresponding to values  $x_1, \dots, x_n$  of the independent variable  $x$ , which constitute a set  $X$  with a partial order  $\prec$  (i.e., a reflexive, transitive and antisymmetric binary relation on  $X$ ), fit to the  $y_i$  a best function  $y = f(x)$  which is non-decreasing (alias



isotonic) with respect to  $\prec$ . The error norm usually chosen is  $l_2$ : we are seeking an isotonic function  $f$  on  $X$  that minimizes

$$D_2(f) = [\sum_{i=1}^n (y_i - f(x_i))^2]^{\frac{1}{2}}.$$

Algorithms for this problem have received a great deal of attention and a collection of them have been discussed in details in [3, 11, 25]. In the case when  $\prec$  is a total order on  $X$  all of the algorithms work in linear time  $O(n)$  provided  $\prec$  is given. In particular we mention the simple Pool-Adjacent-Violators algorithm introduced by Ayer et al. [1] and popularized by Kruskal [18] under the name Up-and-Down Blocks algorithm. It has been extended to rooted trees by Thompson [31] with the Minimum Violator algorithm. The Pool-Adjacent-Violators algorithm is also implicit in van Eeden [9], who extended in [10] the procedures to regressions bounded by two given functions.

The situation becomes much more complex if  $\prec$  is a partial order, say  $x_1, \dots, x_n$  are points in the  $d$ -dimensional space and  $x_i \prec x_j$  if and only if  $x_i = (x_i^{(1)}, \dots, x_i^{(d)})$ ,  $x_j = (x_j^{(1)}, \dots, x_j^{(d)})$  and  $x_i^{(k)} \leq x_j^{(k)}$  for each  $1 \leq k \leq d$ . Although some algorithms described in [3] are applicable in the general case, their computational complexity is already exponential; see [8]. A convergent numerical algorithm for  $X \subset \mathbb{R}^2$  has been proposed in [8]. Geometrically, one can formulate the  $l_2$ -regression problem as the computing of the projection of the vector  $y = (y_1, \dots, y_n)$  onto the convex cone  $K$  of the isotonic functions on  $X$ . Since  $K$  is defined by a finite number of constraints, it is polyhedral. Therefore, one can use any algorithm for solving a quadratic optimization problem, especially those established for projecting onto polyhedral cones, as, for example, that presented in [19]. The alternative procedure of Dykstra [7] is also efficient for polyhedral cones. However, several specific algorithms have been developed for regression problems. The reader will find a vast literature on this topic.

Among other criteria, the choice of  $l_2$  as an error norm is due to the connection with special estimates. Usually, the estimates studied in order restricted statistical inference can be expressed by compact "max-min" formulas (of course, being an useful tool in consistency proofs, these formulas are not very appropriate for computing of the estimates). For instance, let  $M(A)$  be the mean of a collection  $A$  of observations taken from a poset  $X$ . Now, a subset  $L$  of  $X$  is called an *upper layer* if  $x_i \in L$  and  $x_i \prec x_j$  imply that  $x_j \in L$ . Then

$$\hat{f}(x_i) = \max_{\{L: x_i \in L\}} \min_{\{L': x_i \notin L'\}} M(L - L')$$

is an isotonic function and could be used as an estimate. It is shown in [3] that  $\hat{f}$  provides the best  $l_2$ -approximation of  $y_1, \dots, y_n$  and can be calculated by a special minimum lower sets algorithm introduced by Brunk et al. [5] (the complexity of the latter is exponential). A deep generalization of this result to all Cauchy mean value functions has been obtained in [28] (a function  $M$  defined on the nonempty subsets of  $X$  is said to be a *Cauchy mean value function* [27, 28] if  $M(A \cup B)$  belongs to the segment  $[M(A), M(B)]$  whenever  $A$  and  $B$  are nonempty and disjoint subsets of  $X$ ). A description of all linear Cauchy mean value functions has been given in [21]. The result from [28] asserts that if  $M$  is such a function and the error measure  $D(f)$  verifies three rather natural conditions, then  $\hat{f}$  given by the abovementioned "min-max" formula minimizes  $D(f)$  subject to the restriction that  $f$  is isotonic on  $X$ . Moreover,  $\hat{f}$  can be computed using a refined version of the minimum lower set algorithm. As is noticed in [28], this result includes all  $l_p$ -regression problems ( $1 \leq p \leq \infty$ ) in their most general form as special cases (however, the modal regression problem [29] does not fit in this framework).

Namely, assume that with each element  $x_i$  of a poset  $(X, \prec)$  is associated a set of numbers  $y_{i1}, \dots, y_{ir_i}$ , corresponding, for example, to a sample from the  $i$ th distribution. We are looking for an isotonic function  $f$  on  $X$  (i.e.,  $x_i \prec x_j$  implies  $f(x_i) \leq f(x_j)$ ) that minimizes

$$D_p(f) = \left[ \sum_{i=1}^n \sum_{l=1}^{r_i} (y_{il} - f(x_i))^p \right]^{\frac{1}{p}}.$$

If  $p = 1$  we obtain the isotonic median regression problem that corresponds to the  $l_1$ -error norm:

$$D_1(f) = \sum_{i=1}^n \sum_{l=1}^{r_i} |y_{il} - f(x_i)|.$$

For  $X \subset \mathbb{R}$  this problem have been investigated in [26, 24] (unfortunately, the algorithm presented in [24] contains a serious gap, since its two steps do not cover all possible cases). The minimum lower set algorithm from [28], acting on the collection of upper layers, has an exponential complexity already if this collection is of exponential cardinality. This is the case of rather simple partial orders as rooted trees or series-parallel partial orders. To our knowledge there are no polynomial time algorithms to find a best isotonic  $l_p$ -regression function ( $p < \infty$ ) in the general case of a partial order or even in case when  $X \subset \mathbb{R}^d$ ,  $d \geq 2$ . It will be interesting and important to know *what instances of this problem are NP-complete algorithmical problems*.

## 2 Isotonic $l_p$ -regression problem for rooted trees

This section is devoted to the isotonic  $l_p$ -regression problem ( $1 \leq p \leq \infty$ ) for partial orders whose covering graphs are rooted trees. We propose a simple extension of the maximum (minimum) violator algorithm, originally established for the  $l_2$ -criterion. Let us consider a rooted tree, the vertices of which are the elements  $x_1, \dots, x_n$  of a set  $X$ . The tree is here oriented from the leaves to the root. To every vertex  $x_i$ , a sample  $y_{i1}, \dots, y_{ir_i}$  is preassigned. We consider the  $l_p$ -regression problem defined in the introduction: minimize  $D_p(f)$  subject to the constraint that  $f$  is isotonic. The  $y_{il}$  define a vector of  $\mathbb{R}^r$ , where  $r = \sum_{i=1}^n r_i$ . Noting  $\theta_i = f(x_i)$  and duplicating  $r_i$  times the  $\theta_i$ , we get a current vector  $\theta \in \mathbb{R}^r$ , and the minimization problem may be written as: minimize  $\|y - \theta\|_p$ , with equality and inequality constraints over  $\theta$ . Such a problem corresponds to the projection onto a closed polyhedral cone, with respect to a norm  $l_p$ . That provides the existence of a solution.

For every sample of real numbers  $A = \{u_1, \dots, u_k\}$ , let us consider the problem:  $\min_x [\sum_{j=1}^k |x - u_j|^p]^{\frac{1}{p}}$ . It is well-known that such a problem admits a solution. This is the mean for  $p = 2$ , the midrange for  $p = \infty$  and a median point for  $p = 1$ . Every solution lies between  $\min u_j$  and  $\max u_j$ , and due to the convexity, the set of solutions is a closed interval. We denote it by  $M(A) = [a, b]$ . Moreover, it is easy to see that  $[\sum_{j=1}^k |x - u_j|^p]^{\frac{1}{p}}$  is strictly decreasing when  $x$  varies from  $-\infty$  to  $a$ , and is strictly increasing when  $x$  varies from  $b$  to  $+\infty$ . Again, given two samples  $A$  and  $A'$ , we denote by  $A + A'$  their amalgamation. Let  $M(A) = [a, b]$ ,  $M(A') = [a', b']$ ,  $M(A + A') = [c, d]$ . Then, clearly  $M(A + A') = M(A) \cap M(A')$  provided  $M(A) \cap M(A') \neq \emptyset$ . Moreover, it is easy to prove that  $M$  obeys the Cauchy mean condition: if  $M(A) \cap M(A') = \emptyset$  with, for instance,  $b < a'$ , then  $b \leq c$  and  $d \leq a'$ .

Now we describe our algorithm. At a current step, we have a rooted tree with a partition of  $X$ , say  $V$ , as a set of vertices. To every vertex  $v_i$  of  $V$ , a sample  $A_i$  is assigned, where  $A_i$  stands for the amalgamation of the samples  $\{y_{jl}, l = 1, \dots, r_j\}$  for all  $j$  such that  $x_j \in v_i$ . Initially,  $V$  is the finest partition of  $X$ . For every  $v_i \in V$ , denote  $M(A_i) = [a_i, b_i]$ . If for every edge  $(v_i, v_j)$  of the current tree with  $v_i \prec v_j$  we have  $a_i < a_j$ , then put  $\hat{f}(x_k) = a_i$  for every  $x_k \in v_i$  and stop. Otherwise, find  $v_0 \in V$  with predecessors  $v_1, \dots, v_m$  obeying the following conditions

- (i)  $a_0 \leq a_{\max} := \max_{i=1, \dots, m} a_i$ ;
- (ii) for all  $i = 1, \dots, m$  if  $v_i \prec v_0$  then  $a_i < a_0$ .

Then aggregate  $v_0$  and all  $v_i$  such that  $a_i = a_{\max}$ , and amalgamate the corresponding samples  $A_0$  and  $A_i$ . We get a new partition of  $X$  with

$v_0 := v_0 \cup \{v_i : a_i = a_{max}\}$  and all other subsets of  $V$ , and a new rooted tree, to which we apply the same procedure.

**Proposition 1**  *$\hat{f}$  obtained by this algorithm minimizes the  $l_p$ -criterion function  $D_p(f)$  subject to the restriction that  $f$  is isotonic. The function  $\hat{f}$  can be defined in total  $O(nr)$  number of operations.*

**Proof:** The current step corresponds to the new isotonic  $l_p$ -regression problem:

(P) : minimize  $[\sum_{v_i \in V} \sum_{y \in A_i} |\theta_i - y|^p]^{\frac{1}{p}}$ , with the constraints:  $v_i \prec v_j$  implies  $\theta_i \leq \theta_j$ .

We prove by induction that every solution  $\hat{\theta}$  of the reduced problem furnishes a solution  $\hat{f}$  of the initial problem by letting  $\hat{f}(x_k) = \hat{\theta}_i$  provided  $x_k \in v_i$ . Clearly, that is true at the initial step. Suppose that at a current step, there is  $v_0$  obeying the conditions of the algorithm.

First, we show that there is a solution  $\hat{\theta}$  of (P) verifying  $\hat{\theta}_0 = \hat{\theta}_j$  for some  $j \in \{1, \dots, m\}$ . Suppose there is a solution  $\theta^*$  of (P) verifying:  $\theta_0^* > \theta_i^*$  for all  $i = 1, \dots, m$ . Then  $a_i \leq \theta_i^* < \theta_0^* \leq b_0$  for all  $i = 1, \dots, m$ . Indeed, if for some  $i$ ,  $\theta_i^* < a_i$ , defining  $\theta'$  by  $\theta'_i := \theta_i^* + \delta$  and  $\theta' := \theta^*$  otherwise,  $\theta'$  should satisfy the isotony constraints and should reduce the value of the criterion for  $\delta > 0$  sufficiently small. Similarly, if  $\theta_0^* > b_0$ , defining  $\theta''$  by  $\theta''_0 := \theta_0^* - \delta$  and  $\theta'' := \theta^*$  otherwise,  $\theta''$  should satisfy the constraints and should reduce the value of the criterion for  $\delta > 0$  sufficiently small, establishing our assertion.

Let  $\theta_j^*$  realize  $\max\{\theta_i^* : i = 1, \dots, m\}$ . Then  $a_0 \leq \max a_i \leq \theta_j^* < \theta_0^* \leq b_0$ . Define  $\bar{\theta}$  by  $\bar{\theta}_j := \theta_j^*$  and  $\bar{\theta} = \theta^*$  otherwise. Then clearly  $\bar{\theta}$  satisfies the isotony constraints and preserves the value of the criterion. Thus,  $\bar{\theta}$  is a solution of a problem obtained from (P) by adding another constraint  $\theta_0 = \theta_j$ , which is clearly equivalent to a reduced problem of (P) by aggregating  $v_0$  and  $v_j$ , by amalgamating  $A_0$  and  $A_j$  to  $A'_0$  and by joining the predecessors of  $v_j$  to  $v'_0 := v_0 \cup v_j$ . The hypotheses show that  $a'_0 \leq a_{max}$ , where  $M(A'_0) = [a'_0, b'_0]$ . By induction and previous assertion one can deduce that there is a solution  $\bar{\theta}$  of (P) verifying  $\bar{\theta}_0 = \bar{\theta}_i$  for every  $i$  such that  $a_i = a_{max}$ . Since every solution of the reduced problem defined by the algorithm is a solution of (P) satisfying those new constraints, the induction is proved. Moreover, at the end of the algorithm,  $\hat{\theta}$  defined by  $\hat{\theta}_i := a_i$  is clearly a solution of (P).  $\square$

The condition required in the algorithm for pooling two adjacent vertices indicates that we must start from leaves and proceed in an up-and-down way. However, when we have a total order, the proof shows that we may aggregate any pair of consecutive vertices violating isotony condition, as in

the Pool-Adjacent-Violator algorithm.

### 3 Isotonic $l_\infty$ -regression problem and its variants

In this section we consider the isotonic regression problem with the  $l_\infty$ -error norm (i.e. the well-known uniform or Chebychev measure of error); for algorithmic approaches to similar approximation problems see [4, 15, 14] and the references there. For this problem we present a strikingly simple optimal estimate which can be computed in time proportional to the size of the covering graph of the poset  $(X, \prec)$ . If  $X \subset \mathbb{R}^d$  and  $|X| = n$  the computational complexity to compute this estimate is  $O(dn^2)$ . The result is due to Ubhaya [32, 33] and was rediscovered by one of the authors of this note. Also we consider the  $p$ -isotonic regression problem with the Chebychev norm and show how to reduce it to a graph-theoretical problem. In the particular case  $p = 2$  this allows to present a polynomial algorithm.

#### 3.1 Isotonic $l_\infty$ -regression problem

Let  $X = \{x_1, \dots, x_n\}$  be a set of observation points endowed with a partial order  $\prec$  and let  $y_1, \dots, y_n$  be the corresponding values of the dependent variable  $y$ . The simplest (and the most economical way) to present a partial order is to define its *covering graph*  $G = (X, E)$ : in  $G$  two elements  $x_i$  and  $x_j$  are joined by an arc if  $x_i \prec x_j$  and there is no other element  $x_k$  so that  $x_i \prec x_k \prec x_j$ . The goal of the isotonic regression problem with the  $l_\infty$ -norm is to determine an isotonic function  $f : X \rightarrow \mathbb{R}$  that minimizes the  $l_\infty$ -error

$$D_\infty(f) = \max_{x_i \in X} |y_i - f(x_i)|$$

(in [12] the history of using this and other criteria in estimation procedures has been discussed).

For an element  $x_i \in X$  consider the order ideals

$$L_{\prec}(x_i) = \{x_j \in X : x_j \prec x_i\}, L_{\succ}(x_i) = \{x_j \in X : x_i \prec x_j\}.$$

Let

$$f^*(x_i) = \max\{y_j : x_j \in L_{\prec}(x_i)\}$$

and

$$f_*(x_i) = \min\{y_j : x_j \in L_{\succ}(x_i)\}.$$

The reflexivity of  $\prec$  implies that  $L_{\prec}(x_i) \cap L_{\succ}(x_i) = \{x_i\}$ . In particular,  $f^*(x_i) \geq f_*(x_i)$ .

**Proposition 2** *The function  $\hat{f}(x_i) = \frac{1}{2}(f^*(x_i) + f_*(x_i))$  minimizes  $D_\infty(f)$  subject to the restriction that  $f$  is isotonic. Given the covering graph  $G$ ,*

the values of  $\hat{f}$  can be computed in total time  $O(|E|)$ . If  $X \subset \mathbb{R}^d$ , then the computation of  $\hat{f}$  can be performed in  $O(dn^2)$  time.

**Proof:** Let  $(x_i, x_j)$  be an arbitrary edge of  $G$ , and assume  $x_i \prec x_j$ . Since  $L_{\prec}(x_i) \subset L_{\prec}(x_j)$  and  $L_{\succ}(x_j) \subset L_{\succ}(x_i)$ , we conclude that  $f^*(x_j) \geq f^*(x_i)$  and  $f_*(x_j) \geq f_*(x_i)$ , yielding that  $\hat{f}$  is isotonic on  $X$ .

Suppose by way of contradiction that  $D_{\infty}(g) < D_{\infty}(\hat{f})$  for an isotonic function  $g$ . Let  $\epsilon = D_{\infty}(\hat{f})$  and consider an element  $x_i$  such that  $\epsilon = |\hat{f}(x_i) - y_i|$ . Let  $f^*(x_i) = y_j$  and  $f_*(x_i) = y_k$  for elements  $x_j \in L_{\prec}(x_i)$  and  $x_k \in L_{\succ}(x_i)$ . Suppose without loss of generality that  $y_i$  belongs to the segment  $[y_k, \hat{f}(x_i)]$ . Since  $|g(x_i) - y_i| < \epsilon$  and  $g(x_j) \leq g(x_i)$  we immediately obtain that  $g(x_j) < \hat{f}(x_i)$ . But then  $|g(x_j) - y_j| > y_j - \hat{f}(x_i) > \epsilon$ , contrary to the choice of  $g$ .

The values of  $f^*$  can be computed recursively, starting from the minimal elements of  $X$ . If we know  $f^*$  for all predecessors of  $x_i$ , then  $f^*(x_i)$  is the maximum among the  $y_i$  and  $\max\{f^*(x_j) : (x_j, x_i) \in E\}$ . Analogously, the values of  $f_*$  can be computed recursively starting from the maximal elements of  $X$ . If  $f_*$  is computed for all successors of  $x_i$ , then  $f_*(x_i)$  is the minimum among the  $y_i$  and  $\min\{f_*(x_j) : (x_i, x_j) \in E\}$ . Evidently, this can be done in  $O(|E|)$  time. If  $X \subset \mathbb{R}^d$ , then the covering graph of the resulting poset can be computed in  $O(dn^2)$  time.  $\square$

In Figure 1 we present an example of application of Proposition 2 (the optimal error is  $\epsilon^* = 4$ ).

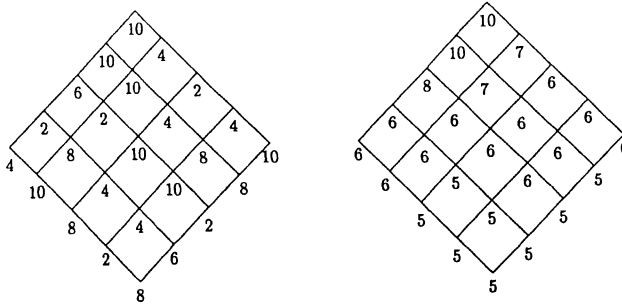


FIGURE 1.

With few efforts one can present an optimal estimate to the general isotonic  $l_{\infty}$ -regression problem. Assume as before that with each element  $x_i$  of  $(X, \prec)$  is associated a set of (distinct) numbers  $y_{i1}, \dots, y_{ir_i}$ , and we wish to find an isotonic function  $f$  on  $X$  that minimizes

$$D_{\infty}(f) = \max_{x_i \in X} \max_{l=1, \dots, r_i} |y_{il} - f(x_i)|.$$

For an element  $x_i \in X$  set

$$f^*(x_i) = \max_{x_j \in L_{\prec}(x_i)} \max_{l=1, \dots, r_j} y_{jl}$$

and

$$f_*(x_i) = \min_{x_j \in L_{\succ}(x_i)} \min_{l=1, \dots, r_j} y_{jl}.$$

We assert that the function  $\hat{f}(x_i) = \frac{1}{2}(f^*(x_i) + f_*(x_i))$  minimizes  $D_\infty(f)$  subject to the restriction that  $f$  is isotonic. For this we extend the partial order  $\prec$  from  $X$  to the multiset  $\{y_{il} : i = 1, \dots, n, l = 1, \dots, r_i\}$  : set  $y_{il} \prec y_{jt}$  if and only if  $x_i \prec x_j$  or  $i = j$  and  $y_{il} \geq y_{jt}$ . Let  $\hat{f}(y_{il})$  be the function defined as in Proposition 2. One can easily note that  $\hat{f}(y_{il}) = \hat{f}(x_i)$  for all  $y_{il}$  ( $l = 1, \dots, r_i$ ). From this and Proposition 2 we deduce that  $\hat{f}(x_i)$  minimizes  $D_\infty(f)$ . The values of  $\hat{f}$  can be computed in  $O(|E| + \sum_{i=1}^n r_i)$  number of operations.

### 3.2 $p$ -Isotonic $l_\infty$ -regression problem

In some recent papers [22, 2, 20] new generalizations of the classical linear regression problem have been given. For example, [20] presents an efficient algorithm for partitioning a planar set  $S = \{s_1 = (x_1, y_1), \dots, s_n = (x_n, y_n)\}$  into two parts  $S_1$  and  $S_2$  such that

$$\sum_{s_i \in S_1} (y_i - f_1(x_i))^2 + \sum_{s_i \in S_2} (y_i - f_2(x_i))^2$$

is minimized, where  $f_1$  and  $f_2$  are the regression lines of the sets  $S_1$  and  $S_2$  (the multidimensional case is treated in [16]). Agarwal and Sharir [2] presented an algorithm with complexity  $O(n^2 \log^5 n)$  for solving a similar problem, replacing the  $l_2$ -criterion function by the  $l_\infty$ -error function. Namely, they are searching a bipartition of a planar set  $S$  such that their maximum width is as small as possible. Recall that the width of a set is the smallest distance between a pair of parallel supporting lines. Equivalently, it is necessary to find two linear functions  $f_1$  and  $f_2$ , such that  $\max_{s_j \in S} \{\min_{i=1,2} |y_j - f_i(x_j)|\}$  is minimized (one can formulate this problem for  $p$ -partitions as is done in [22]). For isotonic regressions, this leads us to the following general formulation.

As before let  $X = \{x_1, \dots, x_n\}$  be a set with a partial order  $\prec$ , and let  $y_1, \dots, y_n$  be the corresponding values of the variable  $y$ . We wish to find a partition  $X_1, \dots, X_p$  of  $X$  and the isotonic functions  $f_1, \dots, f_p$  on  $X_1, \dots, X_p$ , respectively, such that the  $l_\infty$ -error

$$D_\infty(f_1, \dots, f_p) = \max\{\max_{x_i \in X_1} |y_i - f_1(x_i)|, \dots, \max_{x_i \in X_p} |y_i - f_p(x_i)|\}$$

is minimized. Below we will show how to reduce this problem to a special graph-theoretic problem and how to solve it efficiently for  $p = 2$ . Define a symmetric matrix  $D = (d_{ij})$ , where  $d_{ij} = \frac{1}{2}(y_i - y_j)$  if  $x_i \prec x_j$  and  $y_i \geq y_j$ , and  $d_{ij} = 0$  otherwise. Let  $\epsilon^*$  be the minimum of the function  $D_\infty(f_1, \dots, f_p)$ . From Proposition 1 we obtain the following result.

**Lemma 1**  $\epsilon^*$  is an element of the matrix  $D$ .

To find an optimal partition of  $X$  with respect to the criterion function  $D_\infty(f_1, \dots, f_p)$  we proceed as follows (the idea is borrowed from the methods of solving center location problems; see for example [17, 30]). We sort the elements of the matrix  $D$  in the increasing order and search the obtained list for the minimum value which is feasible in the following sense. A value  $\epsilon$  is *feasible* if there is a  $p$ -partition  $X_1, \dots, X_p$  of  $X$  and the isotonic functions  $f_1, \dots, f_p$  on  $X_1, \dots, X_p$ , respectively, such that  $D_\infty(f_1, \dots, f_p) \leq \epsilon$ . To decide if a value  $\epsilon \in D$  is feasible we define a new graph  $\Gamma_\epsilon$ . The vertices of  $\Gamma_\epsilon$  are the elements of  $X$ , and two vertices  $x_i$  and  $x_j$  are adjacent in  $\Gamma_\epsilon$  if and only if either  $x_i$  and  $x_j$  are incomparable or  $x_i \prec x_j$  are comparable and  $y_i - y_j \leq 2 \cdot \epsilon$ . A *clique* of  $\Gamma_\epsilon$  is a subset of pairwise adjacent vertices.

**Lemma 2**  $\epsilon$  is a feasible value if and only if the vertices of  $\Gamma_\epsilon$  can be covered with at most  $p$  cliques.

**Proof:** First, assume that  $D_\infty(f_1, \dots, f_p) \leq \epsilon$  for isotonic functions  $f_1, \dots, f_p$  defined on classes  $X_1, \dots, X_p$  of a partition of  $X$ . We assert that each  $X_k$  is a clique of the graph  $\Gamma_\epsilon$ . Assume the contrary, i.e.  $y_i - y_j > 2 \cdot \epsilon$  for some  $x_i, x_j \in X_k$ ,  $x_i \prec x_j$ . We can suppose without loss of generality that  $y_j$  is the smallest value in  $\{y_s : x_s \in L_{\succ}(x_i)\} \cap X_k$ . Additionally, we can assume that  $f_k$  is defined as in Proposition 2. By this result  $f_k(x_i) = \frac{1}{2}(y_j + y_t)$ , where  $y_t$  is the largest value in  $\{y_s : x_s \in L_{\prec}(x_i)\} \cap X_k$ . Since  $|y_i - f_k(x_i)| \leq \epsilon$  and  $f_k(x_i) \leq f_k(x_j)$ , from  $y_i - y_j > 2 \cdot \epsilon$  one can easily deduce that  $|y_j - f_k(x_j)| > \epsilon$ , contrary to feasibility of  $\epsilon$ . Therefore, if  $\epsilon$  is feasible, then  $X_1, \dots, X_p$  are cliques of the graph  $\Gamma_\epsilon$ .

Conversely, let  $X_1, \dots, X_{p'}$  be a covering of the vertices of  $\Gamma_\epsilon$  with  $p'$  cliques ( $p' \leq p$ ). Let  $\hat{f}_k$  be the isotonic function on  $X_k$  defined in Proposition 2. Pick an arbitrary element  $x_i \in X_k$ . Then  $\hat{f}_k(x_i) = \frac{1}{2}(y_j + y_t)$  where  $y_j$  is the smallest value in  $\{y_s : x_s \in L_{\succ}(x_i)\} \cap X_k$  and  $y_t$  is the largest value in  $\{y_s : x_s \in L_{\prec}(x_i)\} \cap X_k$ . Since  $x_t \prec x_i \prec x_j$  and  $x_t, x_j \in X_t$  we deduce that  $0 \leq y_j - y_t \leq 2 \cdot \epsilon$ . Since  $y_i \in [y_t, y_j]$  we immediately obtain that  $|y_i - \hat{f}_k(x_i)| \leq \epsilon$ . Therefore  $D_\infty(\hat{f}_1, \dots, \hat{f}_{p'}) \leq \epsilon$ , i.e.  $\epsilon$  is feasible.  $\square$

The problem of covering of a graph with a given number of cliques is known to be *NP*-complete [13], however in the particular case  $p = 2$  it can



be easily solved. Indeed, a graph can be covered with two cliques if and only if its complement  $\bar{\Gamma}_\epsilon$  of  $\Gamma_\epsilon$  one can simply use the breadth-first search; see [13] for details. We can construct directly  $\bar{\Gamma}_\epsilon$ : two elements  $x_i, x_j \in X$  are adjacent in  $\bar{\Gamma}_\epsilon$  if and only if  $x_i \prec x_j$  and  $y_i - y_j > 2 \cdot \epsilon$ . Therefore, to solve the initial regression problem by Lemma 1 we must find the smallest feasible value in  $D$ . We use the binary search in the ordered matrix  $D$ . Namely, we start from a median  $\epsilon$  of this list. We construct the graph  $\Gamma_\epsilon$  and check if this graph has a covering with  $p$  cliques. If the answer is "yes" we continue the search in the first half of the list (removing the second sublist from further considerations). Otherwise, if the answer is "not", then we remove the first half and continue the search in the sublist of  $D$  containing the elements larger than  $\epsilon$ . In the current list we take a median element as a current  $\epsilon$  and check if it is feasible. We continue the procedure, until we arrive at a list containing only one element  $\epsilon^*$ . This is the optimal error for the formulated regression problem, while any covering  $X_1, \dots, X_p$  of  $\Gamma_{\epsilon^*}$  with at most  $p$  cliques and the isotonic functions  $\hat{f}_1, \dots, \hat{f}_p$  defined on  $X_1, \dots, X_p$  according to Proposition 2, represent the optimal solution. To find it, we must perform  $O(\log n)$  feasibility tests (namely,  $\log n^2$  such tests) in the sorted matrix  $D$  (to order the elements of  $D$  we need  $O(n^2 \log n)$  operations). The graph  $\Gamma_\epsilon$  can be constructed in  $O(n^2)$  time. If  $p = 2$  within the same time bounds one can decide if  $\bar{\Gamma}_\epsilon$  is bipartite. Therefore, the whole complexity of the algorithm for  $p = 2$  is  $O(n^2 \log n)$ .

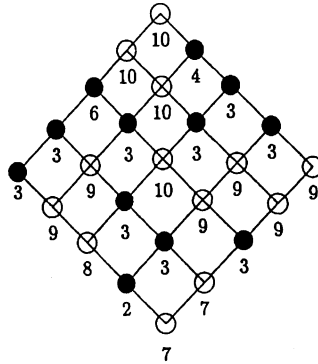


FIGURE 2.

**Proposition 3** *For  $p = 2$  the optimal bipartition of  $X$  with respect to the  $l_\infty$ -criterion function can be constructed in  $O(n^2 \log n)$  number of operations.*

In Figure 2 we present an optimal bipartition of the poset from Figure

1. Note that the optimal error is  $\epsilon^* = 1$ .

### 3.3 Isotonic $l_\infty$ -regression problem with a given number of values

Some papers [4, 15, 14] consider the following approximation problem: given an integer  $p$  and  $(x_1, y_1), \dots, (x_n, y_n)$  in  $\mathbb{R}^2$  with  $x_1 < \dots < x_n$  find a piecewise-linear function  $f$  with at most  $p$  links such that  $\max_{i=1, \dots, n} |f(x_i) - y_i|$  is minimized. Efficient algorithms for solving this problem are presented in [15, 14], for motivation see [4, 15]. If instead of piecewise-linear functions we consider stepwise functions with a fixed number of steps we obtain a particular case of the rectilinear center trajectory problem investigated in [6] (for the latter problem [6] presents an algorithm with the complexity  $O(n^p)$ ). In this section we consider the following regression problem: consider the numbers  $x_1 < \dots, x_n$ , and, assume that with each  $x_i$  is associated a set of (distinct) numbers  $y_{i1}, \dots, y_{ir_i}$ . Given an integer  $p$  we wish to find an isotonic stepwise function  $f$  that minimizes

$$D_\infty(f) = \max_{x_i \in X} \max_{l=1, \dots, r_i} |y_{il} - f(x_i)|$$

subject to the restriction that  $f$  takes at most  $p$  distinct values. Let  $a_i = \min_{l=1, \dots, r_i} y_{il}$  and  $b_i = \max_{l=1, \dots, r_i} y_{il}$ . The key observation is that, as in the previous section, the optimal error  $\epsilon^*$  of  $D_\infty$  is an element of the matrix  $D = (d_{ij})$ , where  $d_{ij} = |b_i - a_j|$ . Therefore we can use a binary search in the ordered list of the elements of  $D$ . With a current  $\epsilon \in D$  we must answer the following question: "There is an isotonic function  $f$  with at most  $p$  steps such that  $D_\infty(f) \leq \epsilon$ ?" To perform this test we proceed as follows. For a given  $x_i$  denote by  $S_i$  the intersection of the segments  $[a_i - \epsilon, a_i + \epsilon]$  and  $[b_i - \epsilon, b_i + \epsilon]$ . We sweep the list  $x_1, \dots, x_n$  from left to right. We need three parameters  $S$ ,  $q$  and  $\mathcal{S}$  whose meaning shall become clear immediately. Initially, let  $S := S_1$ ,  $q := 0$  and  $\mathcal{S} = (-\infty, x_1]$ . At point  $x_i$  we do the following. Find  $S \cap S_i$ . If this intersection is nonempty, then set  $S := S \cap S_i$ ,  $\mathcal{S} = \mathcal{S} \cup (x_{i-1}, x_i]$  and go to the point  $x_{i+1}$ . Otherwise, if  $S \cap S_i = \emptyset$ , then for all  $x \in \mathcal{S}$  define  $f(x) := s$ , where  $s$  is an arbitrary value from the segment  $S$ . If  $x_i < s$ , then stop: the test has a negative answer. Otherwise, set  $S := S_i$ ,  $q := q + 1$ ,  $\mathcal{S} = (x_{i-1}, x_i]$  and consider the next point  $x_{i+1}$  (of course, if  $i = n$  we simply put  $f(x_i) = y_i$  and finish the procedure). After  $n$  steps we return answer "yes" if  $q \leq p$  and the answer "no", otherwise. The complexity of this procedure is  $O(n)$ . The proof of correctness is straightforward. To find an optimal isotonic function  $f$  with at most  $p$  values we must perform  $O(\log n)$  feasibility tests in the ordered matrix  $D$ . If we simply sort the matrix  $D$ , the total complexity of the algorithm will be  $O(n^2 \log n)$  (actually, this is the time to sort  $D$ ,

because the complexity of testing is only  $O(n \log n)$ . We can improve the whole complexity of our algorithm. Instead of constructing and sorting the matrix  $D$ , we can use the selection algorithm of [23]. It presents an  $O(n \log^2 n)$  time algorithm for computing the  $k$ th largest element in the set of all simple paths in a tree. One can view the sorted list of numbers  $\{a_i, b_i : i = 1, \dots, n\}$  as a path; therefore, we can apply the algorithm from [23]  $O(\log n)$  times, leading us to an algorithm with the total complexity  $O(n \log^3 n)$ .

**Proposition 4** *Given a total order  $x_1 < \dots < x_n$  and an integer  $p > 0$  an isotonic function  $f$  minimizing the  $l_\infty$ -criterion function subject to the restriction that  $f$  has at most  $p$  distinct values can be constructed in  $O(n \log n^3)$  number of operations.*

Most likely, using the parametric search as in [2, 14] one can solve this problem more efficiently. We leave open the question whether a similar problem for all partial orders is  $NP$ -complete. Finally note that within the same time bounds we can solve the problem of approximating with a stepwise function with at most  $p$  distinct values. Again, the optimal  $l_\infty$ -error is an element of the matrix  $D$ . We can apply the same test, but in case  $S \cap S_i = \emptyset$  it is not necessary to check whether  $x_i < s$ .

## References

- [1] Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26**, 641–647.
- [2] Agarwal, P. and Sharir, M., Planar (1994). Geometric location problems. *Algorithmica* **11**, 185–195.
- [3] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions*. New York: Wiley.
- [4] Bellman, R.E. and Roth R.S. (1969). Curve fitting by segmented straight lines. *J. Am. Statist. Assoc.* **64**, 1079–1084.
- [5] Brunk, H.D., Ewing, G.M. and Utz, W.R. (1957). Minimizing integrals in certain classes of monotone functions. *Pacif. J. Math.* **7**, 833–847.
- [6] Diaz-Banez, J.M. and Mesa, J.A. (1995). An algorithm for a rectilinear center trajectory. Preprint Universidad de Sevilla.
- [7] Dykstra, R.L. (1983). An algorithm for restricted least squares regressions. *J. Am. Statist. Assoc.* **78**, 837–842.

- [8] Dykstra, R.L. and Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *Ann. Statist.* **10**, 708–716.
- [9] Eeden, C. van (1956). Maximum likelihood estimation of ordered probabilities. *Proc. Kon. Nederl. Akad. Wetensch. A* **59**, 444–455.
- [10] Eeden, C. van (1958). Testing and estimating ordered parameters of probability distributions. Doctoral Dissertation, University of Amsterdam.
- [11] Eeden, C. van. Estimation in restricted parameter spaces. Some history and some recent developments. Technical report 163, Dept. Stat. University of British Columbia, Vancouver, Canada (to appear in *CWI Quarterly*).
- [12] Farebrother, R.W. (1987). The historical development of the  $L_1$  and  $L_\infty$  estimation procedures. In *Statistical Data Analysis Based on the  $L_1$ -Norm and Related Methods*, Ed. Y. Dodge, pp. 37–63.
- [13] Golumbic M.C. (1980). *Algorithmic Graph Theory and Perfect Graphs*. New York: Academic Press.
- [14] Goodrich, M.T. (1995). Efficient piecewise-linear function approximation using the uniform metric. *Discrete Comput. Geom.* **14**, 445–462.
- [15] Hakimi, S.L. and Schmeichel, E.F. (1991). Fitting polygonal functions to a set of points in the plane. *CVGIP: Graph. Mod. Image Proc.* **53**, 132–136.
- [16] Hayes, A.C. and Larman, D.G. (1985). Least squares procedures for problems involving more than one hyperplane. In *Discrete Geometry and Convexity*, Eds. J.E. Goodman, E. Lutwak, J. Malkevitch and R. Pollack, pp. 335–363.
- [17] Kariv, O. and Hakimi, S.L. (1979). An algorithmic approach to network location problems, I,II. *SIAM J. Appl. Math.* **37**, 513–537 and 538–563.
- [18] Kruskal, J.B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* **29**, 115–129.
- [19] Lawson, C. and Hanson, R.J. (1974). *Solving Least Squares Problems*. Prentice-Hall.
- [20] Lenstra, A.K., Lenstra, J.K., Rinnooy Kan A.H.G. and Wansbeek, T.J. (1981). Two lines least squares. *Math. Centrum*. Preprint.
- [21] Leurgans, S. (1981). The Cauchy mean value property and linear functions of order statistics. *Ann. Statist.* **9**, 905–908.
- [22] Megiddo, N. and Supowit, K. (1984). On the complexity of some common geometric location problems. *SIAM J. Comput.* **13**, 182–196.
- [23] Megiddo, N., Tamir, A., Zemel, E. and Chandrasekaran, R. (1981). An  $O(n \log^2 n)$  algorithm for the  $k$ th longest path in a tree with application to location problems. *SIAM J. Comput.* **10**, 328–337.

- [24] Menéndez, J.A. and Salvador, B. (1987). An algorithm for isotonic median regression. *Comput. Statist. Data Anal.* **5**, 399–406.
- [25] Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. New York: Wiley
- [26] Robertson, T. and Wright, F.T. (1973). Multiple isotonic median regression. *Ann. Statist.* **1**, 422–432.
- [27] Robertson, T. and Wright, F.T. (1974). A norm reducing property for isotonized Cauchy mean value functions. *Ann. Statist.* **6**, 1302–1307.
- [28] Robertson, T. and Wright, F.T. (1980). Algorithms in order restricted statistical inference and the Cauchy mean value property. *Ann. Statist.* **8**, 645–651.
- [29] Sager, T.W. and Thisted, R.A. (1982). Isotonic modal regression. *Ann. Statist.* **10**, 692–707.
- [30] Tamir, A. (1988). Improved complexity bounds for center location problems on networks by using dynamic data structures. *SIAM J. Discr. Math.* **1**, 377–396.
- [31] Thompson, W.A. Jr. (1962). The problem of negative estimates of variance components. *Ann. Math. Statist.* **33**, 273–289.
- [32] Ubhaya, V.A. (1974). Isotone optimization, I,II. *J. Approxim. Th.* **12**, 146–159 and 315–331.
- [33] Ubhaya, V.A. (1974). Approximation by isotone functions—a min–max form of best approximation. *Notices Amer. Math. Soc.* **21**, A–307.

# Notes on the early history of elemental set methods

**Richard William Farebrother**

*Victoria University of Manchester, UK*

*Abstract:* In this paper we outline the early history of traditional estimation procedures which are based on the use of elemental sets. There are two distinct classes of such procedures associated with the minimum values of the sum of absolute errors and the largest absolute error criteria respectively. As a matter of historical necessity, our study will concentrate on estimation procedures of the first type. However we shall also discuss some recent work on the least median of squared errors procedure which in principle involves elemental sets of the second type.

*Key words:* Least median of squared errors, least sum of absolute errors, least sum of squared errors, linear programming, minimax absolute error.

AMS subject classification: 01A50, 01A55, 62J05.

## 1 Introduction

In his study of the practical value of elemental set approximations to robust estimation procedures, Hawkins (1993, p.580) has summarised the history of such methods in the following terms:

Elemental set methods have their origins in the single-predictor proposal by Theil (1950). The extension of the idea to handling outlier problems in multiple regression was made independently by Rousseeuw (1984) and by Hawkins, Bradu and Kass (1984). They also arise naturally in the expression of the OLS multiple regression in terms of weighted U-statistics (as sketched in the technical appendix to Hawkins *et al.*, 1984).

Although this brief statement may well have been sufficient for the purposes of Hawkins's paper, it cannot be regarded as an adequate summary of the history of elemental set methods as it fails to mention that meth-

ods of this type had been used to estimate the parameters of linear and nonlinear models with two or more predictors for more than 240 years, or that the elemental set characterisation of the ordinary least squares (OLS) estimator has been known for more than 150 years. The purpose of the present paper is to provide a more comprehensive survey of the history of this subject. We shall also take the opportunity to mention some recent work on the least median of squared errors procedure.

## 2 Nature of the fitting problem

Some readers of this paper may be prompted to further their historical studies by consulting some of the original sources cited in it. Such readers will soon discover that the traditional notation and nomenclature of the Calculus of Observations is quite distinct from that of modern Mathematical Statistics. To help the interested reader through this difficulty we shall employ a variant of this traditional usage in our history.

In traditional notation the familiar curve or surface fitting problem may be expressed in the following terms: We are given a linear or nonlinear function  $f(\cdot)$  which is characterised by a set of  $p$  observed (variable) quantities  $a, b, c, \text{etc.}$  and a set of  $q$  unobserved fixed quantities (we would call them parameters)  $x, y, z, \text{etc.}$  We suppose that the variable quantities  $a, b, c, \text{etc.}$  are observed without error but that the value of the function  $f(\cdot)$  is subject to an additive error. Denoting the observed value of the function by  $m$  and the corresponding value of the additive error by  $v$ , we find that we have a system of  $n$  equations:

$$m_i = f(a_i, b_i, c_i, \text{etc.}; x, y, z, \text{etc.}) + v_i \quad i = 1, 2, \dots, n$$

which describes the relationship between the observed values of the variable quantities and the corresponding observed values of the function. In this context we have to choose values for the  $q$  unknown quantities  $x, y, z, \text{etc.}$  in such a way that the observed errors (we would say residuals)  $v_1, v_2, \dots, v_n$  are as small as possible in some sense.

Although this general statement of the problem admits the possibility of nonlinear functions, we shall be largely concerned with functions  $f(\cdot)$  which are linear in the unknown constants. The major exceptions to this rule being found in the final paragraph of Section 3.

In passing we note that this traditional notation does not distinguish between the true and fitted values of the errors or between the true and fitted values of the unknown constants. This imprecision is of no immediate consequence as we shall only be concerned with fitted values in the remainder of this paper.

### 3 Elemental set methods

The most obvious solution to the general fitting problem outlined in Section 2 is to discard  $n - q$  of the available equations and to use the remaining  $q$  equations to determine a set of values for the  $q$  unknown constants in such a way that these  $q$  equations are exactly satisfied with zero errors. This fitting procedure is variously known as the method of selected points, the subset selection method, and the method of elemental sets.

Now there are  ${}^nC_q$  distinct ways of choosing  $q$  equations from a set of  $n$  equations, so that practitioners have either to choose a single set of  $q$  equations at random or according to some specified rule, or they have to select a greater number of sets and attempt to reconcile the discordant results obtained from their selection.

In the earliest period of enquiry in this area, scientists employed a variant of the first procedure in which the selection criterion is obscure, if not entirely hidden. Without making any attempt to explain their reasons, these authors arranged that there should be as many equations as were required for the problem to have a unique solution. For example, in his analysis of the problem of determining the height of a tree on a remote hillside, Liu Hui (third century) assumed that the surveyor had taken observations on the elevation of the top of the tree from each of two locations in a horizontal reference plane and a third observation on the elevation of the base of the tree from one of these locations. These three observations together with the known horizontal distance between the two observation sites are sufficient to determine the three unknowns of the problem, see Li and Du (1987, pp.76–78) for details. However it is not explained why the surveyor should not have observed both elevations from both locations, or what was to be done if he had. It is intriguing to speculate on how Liu Hui would have responded if he had been challenged on this point.

By the middle of the eighteenth century, this implicit choice of a single set of  $q$  equations had been replaced by a more explicit procedure. Mayer (1750, p.150), for example, suggested that one should choose a set of  $q$  equations that are typical of the  $n$  given equations. In the particular case of his determination of the position of the lunar crater Manilius, he obtained a system of  $n = 27$  equations in  $q = 3$  unknowns of the form

$$m_i = x + y \sin(\theta_i) + z \cos(\theta_i) + v_i \quad i = 1, 2, \dots, n$$

and suggested that the  $q = 3$  typical equations should be chosen in such a way that two of the angles differ from the third by 90 and 180 degrees respectively. However, having advanced this basic suggestion, he observed that one needs more than a single set of typical equations if one wishes to



check the accuracy of the original solution. In the limiting case one thus obtains a total of  $[n/q]$  distinct solutions from disjoint subsets of the data which then have to be reconciled with one another.

Unfortunately Mayer did not discuss this technique in sufficient detail for us to be sure how many disjoint subsets of  $q$  equations he would have used, or how he would have attempted to reconcile the corresponding discordant results. Nevertheless his lack of precision in this regard is entirely comprehensible as his exposition of this topic was as a brief aside to his main purpose of introducing a new fitting procedure known as the method of averages.

A few years later, Boscovich addressed a similar problem relating to the ellipsoidal figure of the Earth in a similar way. In his contribution to Maire and Boscovich (1755) and again in Boscovich (1757), Boscovich was concerned with the solution of a system of  $n = 5$  linear equations in  $q = 2$  unknowns. In his first approach to this problem he evaluated all  ${}^5C_2 = 10$  pairwise determinations of the unknown constants before taking an unweighted arithmetic mean. He was not satisfied with the result and tentatively suggested a variant which discards the two determinations with the smallest denominators before again taking an unweighted average of the remaining eight values.

Boscovich was far from satisfied with the results he obtained from either variant of this procedure and, like Mayer before him, resolved the impasse by proposing an alternative fitting procedure. Boscovich's alternative procedure is to be found in his scientific notes to a poem in Latin hexameters by Stay (1760, pp.420–425), see Farebrother (1993). It chooses values for the unknown constants in such a way as to minimise the sum of the absolute values of the observed errors subject to the condition that the corresponding sum of the signed errors is zero. The relationship between this procedure and the method of elemental sets will be outlined in the following section.

Finally, in this connection, we must observe that the method of elemental sets was not entirely supplanted by more advanced methods (notably the method of least squares) until well into the present century. This statement is particularly true of nonlinear problems as Pearson (1902, p.298) mentions the possible use of the method of elemental sets to fit Makeham's law (a generalisation of the Gompertz function) to actuarial data, and Yule (1925, pp.49–50) notes that this method produces acceptable results when fitting a logistic function to sufficiently smooth demographic data. Indeed, the more elementary statistical textbooks of the 1950s and 1960s still recommended the method of elemental sets for the latter purpose, see Croxton and Cowden (1939; 1955, p.215).

## 4 Elemental set characterisations

Boscovich provided a geometrical algorithm as an integral part of his solution procedure. This algorithm was subsequently given an analytical form by Laplace (1793), see Farebrother (1993), Sheynin (1973), or Stigler (1986) for details. Some years later Gauss (1809, sec.186) generalised Boscovich's optimality criterion to any number of unknowns and suggested that the adding-up constraint could be deleted. (In passing we note that this adding-up constraint has no direct connection with the method of least squares which was developed by Gauss and Legendre some years after Boscovich's death.)

In his discussion of the unconstrained least sum of absolute errors problem, Gauss (1809, sec.186) notes that the optimal solution to this problem is characterised by a set of  $q$  zero errors and that the other  $n - q$  errors only help to determine this optimal set. He gives no justification for this result but a proof which would have been accessible to Gauss and his contemporaries has been suggested by Waterhouse (1990).

An explicit characterisation of the solution to the least squares problem as a weighted sum of elemental set determinations was established by Jacobi (1841). See Sheynin (1973) for an excellent description of this result which closely follows Jacobi's own derivation. This result was subsequently rediscovered by Glaisher (1879), Subrahmanyam (1972), Hawkins, Bradu and Kass (1984), and Ben-Tal and Teboulle (1990) amongst others. The long interlude between the publication of the papers by Glaisher and Subrahmanyam would seem to be due to the intervention of a clear statement of Jacobi's result in the popular textbook by Whittaker and Robinson (1924; 1944, pp.251–252).

This explicit characterisation of the solution to the least squares problem may be generalised to a weighted sum of least squares determinations from sets of  $m > q$  equations, see Sheynin (1993) and Wu (1986) for details. As a further generalisation of this result, Ben-Tal and Teboulle (1990) have shown that, for all members of a class of strictly isotone functions which includes the weighted sum of the  $k$ th powers of the absolute errors (for some fixed finite positive value of  $k$ ), every set of values for the unknown constants which minimises the chosen function of the errors will lie within the convex hull of the elemental set determinations. In addition they have shown that, for all members of a class of isotone (but not strictly isotone) functions including the largest absolute error and the median absolute error functions, at least one of the sets of values for the unknown constants which minimise the chosen function of the errors must lie within the convex hull of the elemental set determinations of these values.

Thus, although the solutions to the least sum of absolute errors ( $k = 1$ ) and the least sum of squared errors ( $k = 2$ ) problems must lie in the convex hull of the elemental set determinations, the solutions of the minimax absolute errors and least median of (squared) absolute errors problems will necessarily be members of the convex hull only if these solutions are unique.

Further, although all the solutions of the weighted sum of  $k$ th powers (or other strictly isotone) problem must lie in a convex set for all sets of weights, the set of all such solutions need not itself be convex. Gilstein and Leamer (1983) have given a precise characterisation of the nonconvex set of solutions to the weighted least squares problem.

## 5 Minimax absolute error criterion

The class of (Gaussian) elemental set methods discussed in Sections 3 and 4 may be associated with the optimal value of the sum of absolute errors criterion. A second class of (Laplacian) elemental set methods may similarly be associated with the optimal value of the maximum absolute error criterion. As its name implies, the minimax absolute error procedure chooses values for the unknown constants in such a way as to minimise the largest in absolute value of the  $n$  observed errors. This fitting procedure was first discussed by Laplace (1786). Given a set of  $n$  linear equations in  $q = 2$  unknowns, Laplace arbitrarily selected a set of  $r = q + 1$  equations. Using any  $q$  of these equations to eliminate the  $q$  unknowns from the remaining equation, he obtained a single (reduced) equation with a linear function of the  $r$  observed errors on one side and a nonnegative constant on the other. Without further explanation, he asserted that the largest in absolute value of these  $r$  errors is minimised when all  $r$  errors take the same absolute value and their signs are given by the signs of the corresponding coefficients in the single reduced equation.

A determinantal formulation of this procedure was subsequently developed by de la Vallée Poussin (1911). In this alternative formulation of the problem one has to set the  $r$  selected errors proportional to the signs of the cofactors of  $m_1, m_2, \dots, m_r$  in the  $r \times r$  determinant

$$\begin{vmatrix} a_1 & b_1 & c_1 & \dots & m_1 \\ a_2 & b_2 & c_2 & \dots & m_2 \\ \vdots & \vdots & \vdots & & \vdots \\ a_r & b_r & c_r & \dots & m_r \end{vmatrix}$$

where, for notational simplicity, we assume that the trial solution is defined by the first  $r$  equations of the model. The values of the unknown constants and the common absolute value of the  $r$  errors are then obtained by applying

Cramer's Rule to the corresponding system of  $r$  equations in  $r$  unknowns.

Under either of these schemes we have to evaluate all  $n$  errors  $v_1, v_2, \dots, v_n$  and select a new set of  $r$  equations if one or more of these errors is larger in absolute value than the common value of the  $r$  errors in the current set. Laplace (1786) apparently proposed to choose this new set without reference to earlier selections. By contrast, de la Vallée Poussin (1911) introduced an automated selection procedure which Stiefel (1960) subsequently identified with his own (1959) procedure and both as equivalent to a standard simplex implementation of the linear programming dual formulation of the minimax problem.

In this context it is interesting to note that Farebrother (1985) has shown that the minimax absolute error procedure is closely related to the linear programming dual formulation of the least sum of absolute errors problem. Also see Farebrother (1997) for a detailed account of the early history of the minimax absolute error procedure with particular reference to the work of de Prony (1804) and Fourier (1827).

## 6 Median squared absolute error criterion

These results on the minimax absolute error procedure have recently come to prominence as Rousseeuw's (1984) so-called least median of squares procedure actually chooses values for the unknown constants in such a way as to minimise the median or middlemost value of any increasing function of the absolute values of the observed errors. For example, suppose that we wish to minimise the  $h$ th largest squared error where  $h$  is set close to  $n/2$ . Then, conditional on the choice of the  $h - 1$  equations which are to be ignored, the least median of squared errors problem may be expressed in the form of a minimax absolute error problem applied to the  $n - h + 1$  retained equations. The optimal solution to the least median of squared errors problem is thus also characterised by a (Laplacian) elemental set determination of the unknown constants. In principle, we may therefore determine the optimal values of these  $q$  constants by evaluating the median squared error function for a sufficiently large sample of the minimax absolute error determinations of the  $q$  unknowns from sets of  $q + 1$  equations.

However, it is clear that the minimax fitting of a system of  $q + 1$  equations in  $q$  unknowns is vastly more expensive than the direct solution of a set of  $q$  equations in  $q$  unknowns. Rousseeuw and Leroy (1987) have therefore suggested that a sufficiently accurate approximation to the exact least median of squares solution may be obtained by evaluating the median squared error function for a sufficiently large sample of Gaussian elemental set determinations. This conjecture was subsequently confirmed

by Stromberg (1993) whilst Hawkins (1993) has shown that this technique yields satisfactory results for a wide class of robust fitting procedures.

## References

- [1] Ben-Tal, A. and Teboulle, M. (1990). A geometric property of the least squares solution of linear equations. *Linear Algebra and Its Applications* **139** 165–170.
- [2] Boscovich R.J. (1757). De litteraria expeditione per pontificiam ditionem et synopsis amplioris operis . . . , *Bononiensi Scientiarum et Artum Instituto atque Academia Commentarii* **4** 353–396. Reprinted by the Institute for Higher Geodesy, Zagreb, 1961.
- [3] Croxton, F.E. and D.J. Cowden. (1939). *Applied General Statistics*. Prentice–Hall Inc., Englewood Cliffs, New Jersey. Second Edition, Sir Isaac Pitman and Sons, London, 1955.
- [4] De la Vallée Poussin, C.J. (1911). Sur la méthode de l’approximation minimum. *Annales de la Société Scientifique de Bruxelles* **35** 1–16.
- [5] de Prony, G.C.F.M.(1804). *Recherches physico-mathématiques sur la théorie des eaux courantes*. Paris: Imprimerie Impériale.
- [6] Farebrother, R.W. (1985). Unbiased  $L_1$  and  $L_\infty$  estimation. *Commun. Statist. Theory and Method* **14** 1941–1962.
- [7] Farebrother, R.W. (1987). The historical development of  $L_1$  and  $L_2$  estimation procedures 1793–1930. In *Statistical Data Analysis based on the  $L_1$ -norm and Related Methods*, Ed. Y. Dodge, pp. 37–63. Amsterdam: North-Holland Publishing Company.
- [8] Farebrother, R.W. (1992). Geometrical foundations of a class of estimators which minimise sums of Euclidean distances and related quantities. In  *$L_1$ -Statistical Analysis and Related Methods*, Ed. Y. Dodge, pp. 337–349. Amsterdam: North-Holland Publishing Company.
- [9] Farebrother, R.W. (1993). Boscovich’s method for correcting discordant observations. In *R.J. Boscovich: His Life and Scientific Work*, Ed. P. Bursill–Hall, pp. 255–261. Rome: Istituto della Enciclopedia Italiana.
- [10] Farebrother, R.W.(1997). The historical development of the linear minimax absolute residual estimation procedure 1786–1960. *Comput. Statist. Data Anal.* Forthcoming.
- [11] Fourier, J.B.J. (1827). Second Extrait, *Histoires de l’Académie Royale des Sciences de Paris [pour 1824]*, p. 47. Reprinted in his *Oeuvres*, Vol 2. Gautier–Villars et Fils, Paris, 1890, pp. 325 –328.
- [12] Gauss, C.F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*.Hamburg: F. Perthes and I.H. Besser.

- Reprinted in his *Werke*, Vol. 7. F Perthes, Gotha, 1871.
- [13] Gilstein, C.Z. and E.E. Leamer (1983). The set of weighted regression estimates. *J. Am. Statist. Assoc.* **78** 942–948.
  - [14] Glaisher, J.W.L. (1879). On the method of least squares. *Memoirs of the Royal Astronomical Society* **40** 600–614.
  - [15] Hawkins, D.M. (1993). The accuracy of elemental set approximations for regression. *J. Am. Statist. Assoc.* **88** 580–589.
  - [16] Hawkins, D.M., Bradu D. and C.V. Kass (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics* **19** 197–208.
  - [17] Jacobi, C.G.J. (1841). De formatione et proprietatibus determinatum. *Journal für die reine und angewandte Mathematik* **22** 285–318. Reprinted in his *Werke*, Vol.3. Georg Reimer, Berlin, 1884, pp.355–392.
  - [18] Laplace, P.S. (1786). Mémoire sur la figure de la Terre. *Mémoires de l'Académie Royale des Sciences de Paris [pour 1783]*, pp.17–46. Reprinted in his *Oeuvres Complètes*, vol.11. Gauthier–Villars et Fils, Paris, 1895, pp.3–32.
  - [19] Laplace P.S. (1793). Sur quelques points du système du monde. *Mémoires de l'Académie Royale des Sciences de Paris [pour 1789]*, pp.1–87. Reprinted in his *Oeuvres Complètes*. Vol. 11. Gauthier–villars et Fils, Paris, 1895, pp.447–558.
  - [20] Li, Y. and S. Du (1987). *Chinese Mathematics: A Concise History*. Translated by J.N. Crossley and A.W.C. Lun. Oxford University Press, Oxford.
  - [21] Maire, C. and R.J. Boscovich. (1755). *De Litteraria Expeditione per Pontificiam Ditionem*. Rome: N and M Palearini.
  - [22] Mayer, J.T. (1750). Abhandlung über die Umwälzung des Monds um seine Axe. *Kosmographische Nachrichten und Sammlungen [1748]* **1** 52–183.
  - [23] Pearson, K. (1902). On the systematic fitting of curves. *Biometrika* **1** 265–303.
  - [24] Rousseeuw, P.J. (1984). Least median of squares regression. *J. Am. Statist. Assoc.* **79** 871–880.
  - [25] Rousseeuw, P.J. and A.M. Leroy (1987). *Robust Regressions and Outlier Detection*. New York: Wiley.
  - [26] Sheynin, O.B. (1973). R J Boscovich's work on probability. *Archive for History of Exact Sciences* **9** 306–324 and **28**, 173.
  - [27] Sheynin, O.B. (1993). On the history of the principle of least squares. *Archive for History of Exact Sciences* **46** 39–54.

- [28] Stay, B. (1760). Annotated by R.J. Boscovich, *Philosophiae Recentioris... Versibus Traditae...*, Tomus II. Rome: N and M Palearini.
- [29] Stiefel, E.L. (1959). Über diskrete und lineare Tschebyscheff Approximationen. *Numerische Mathematik* **1** 1–28.
- [30] Stiefel, E.L. (1960). Note on Jordan elimination, linear programming and Tchebycheff approximation. *Numerische Mathematik* **2** 1–17.
- [31] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, Massachusetts.
- [32] Stromberg, A.J. (1993). Computing the exact value of the least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM J. Scient. Comput.* **14** 1289–1299.
- [33] Subrahmanyam, M (1972). A property of simple least squares estimates. *Sankhyā B* **34** 355–356.
- [34] Theil, H (1950). A rank-invariant method for linear and polynomial regression analysis. *Koninklijke Nederlandse Akademie van Wetenschappen Proceedings Part A* **53** 386–392, 521–525 and 1397–1412.
- [35] Waterhouse, W.C. (1990). Gauss's first argument for least squares. *Archive for History of Exact Sciences* **41**, 41–52.
- [36] Whittaker, E.T. and G. Robinson (1924). *The Calculus of Observations*. Blackie and Son Ltd., London. Fourth Edition (1944). Reprinted by Dover Publications Inc., New York, 1967.
- [37] Wu, C.F.J. (1986). Jackknife, Bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1295.
- [38] Yule, G.U. (1925). The growth of population and the factors which control it, *J. R. Statist. Soc.* **88** 1–62.

# Computational aspects of censored quantile regression

**Bernd Fitzenberger**

*University of Konstanz, Germany*

*Abstract:* Similar to standard quantile regressions, the censored quantile regression estimate interpolates some data points. This paper discusses the algorithms used in empirical research in light of this interpolation property and compares their performance in a simulation study. The results show that the ranking between algorithms differs depending on the criterion used. The algorithm BRCENS, suggested by the author in the past, performs best in terms of the frequency that the exact censored quantile regression estimates are obtained, it is very competitive in terms of the computation times required and its performance can be noticeably improved, when trying out various starting values. However, BRCENS is not optimal in terms of the root-mean-squared deviation of the coefficient estimates, indicating a high skewness of the distribution of the deviation from the exact estimates. Overall, BRCENS can be recommended for moderate degrees of censoring, whereas all practical algorithms perform quite poorly when a lot of censoring is present.

*Key words:* Censored quantile regression, algorithms, BRCENS, ILPA, NLRQ.

AMS subject classification: 65U05, 62J05.

## 1 Introduction

Censored quantile regressions (CQR's), introduced by Powell (1984, 1986), are an attractive approach to the estimation of the censored regression model with fixed known censoring points.<sup>1</sup> First, compared to Tobit maximum likelihood estimation, cf. Amemiya (1985, chapter 10), CQR's provide consistent estimates under far weaker distributional assumptions. Second,

---

<sup>1</sup>Buchinsky (1997, section 8) and Fitzenberger (1997) provide general guides to CQR's.



CQR's allow us to model the conditional quantiles of the dependent variable as a function of the regressors, cf. Buchinsky (1994), Chamberlain (1994), or Fitzenberger et al. (1995) modelling the conditional wage distribution.

The computation of the CQR estimates involves minimizing a non-differentiable and non-convex distance function. When introducing the censored least absolute deviation (LAD) regression (i.e. the CQR for the special case of the median), Powell (1984) suggested generic optimization routines, which do not take account of the special characteristics of the problem.<sup>2</sup> Paarsch (1984) even resorted to grid search when evaluating the finite sample properties of censored LAD regression. Womersley (1986) analyzed the numerical properties of censored LAD regression for the first time and suggested an algorithm using a finite direct descent method. However, to my knowledge, his code is not available to applied researchers.

More recently, Buchinsky (1994) suggested an Iterative Linear Programming Algorithm (ILPA) involving an iteration of the Barrodale–Roberts–Algorithm (BRA) first developed for standard LAD regression.<sup>3</sup> However, ILPA is not guaranteed to converge and convergence does not guarantee a local minimum of the CQR optimization problem. Building on the characterization of the CQR estimates by the interpolation property presented in the following, Fitzenberger (1994) developed the algorithm BRCENS as an adaptation of the BRA guaranteeing convergence to a local minimum. Koenker and Park (1996) developed a general interior point algorithm for nonlinear quantile regression problems (NLRQ), which they apply to the CQR case. The simulation studies in Fitzenberger (1994, 1997) show that BRCENS performs best in comparison to ILPA and NLRQ in terms of the frequency that the exact global minimum of the CQR optimization problem is obtained. However, the simulation studies show that all algorithms perform quite poorly in the presence of a lot of censoring.

The purpose of this paper is to give an overview on the computational aspects of CQR's and to provide more extensive simulation evidence on the performance of various algorithms currently in use. In the remainder of this section, I present the interpolation property characterizing the CQR estimates. Section 2 describes various algorithms in detail and presents modified versions of the algorithms BRCENS and NLRQ. Section 3 extends the available evidence from simulation studies. Summarizing the simulation results, BRCENS performs quite well in comparison and it can be recommended for moderate degrees of censoring relative to the quantile

---

<sup>2</sup>Following Powell's suggestion, Horowitz and Neumann (1987) present an empirical application based on the Nelder and Mead (1965) algorithm.

<sup>3</sup>Koenker and d'Orey (1987) provide an extension of the BRA to the general quantile regression case. For the latter, the following discussion refers to this extension.

considered. However, in the presence of a lot of censoring all algorithms perform quite poorly.

### 1.1 Censored quantile regression and interpolation property

Introducing some notation, for a sample of size  $N$ , let the dependent variable be the  $N \times 1$  vector,  $y = (y_1, \dots, y_N)$ , the design matrix be the  $N \times k$  matrix  $X = (x_1, \dots, x_N)'$ , with  $x_i = (x_{i,1}, \dots, x_{i,k})$ , the  $N \times 1$  vector of fixed known observation specific censoring values be  $yc = (yc_1, \dots, yc_N)$ , the  $N \times 1$  vector of disturbances be  $\epsilon = (\epsilon_1, \dots, \epsilon_N)'$  and the  $k \times 1$  parameter vector be  $\beta$ .

The following discussion considers a censored regression model with censoring from above. For a given quantile  $\theta \in (0, 1)$ , the CQR estimation problem<sup>4</sup> is to minimize the piecewise linear distance function given by

$$\widehat{\beta}_\theta \in \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^N \operatorname{sgn}_\theta(y_i - \min[x_i'\beta, yc_i]) \cdot (y_i - \min[x_i'\beta, yc_i]) \quad , \quad (1)$$

where the  $\theta$  weighted sign function is given by

$$\operatorname{sgn}_\theta(\epsilon_i) \equiv \theta I(\epsilon_i > 0) - (1 - \theta)I(\epsilon_i < 0)$$

and  $I(\cdot)$  denotes the indicator function. The expression  $x_i'\beta_\theta$  captures the  $\theta$ -quantile of the underlying uncensored dependent variable conditional on  $x_i$ .

Since the CQR distance function (1) is piecewise linear, the CQR minimization problem does not necessarily have a unique solution.<sup>5</sup> Analogous to standard quantile regressions, cf. Koenker and Bassett (1978), the following interpolation property can also be established for CQR's.

**Interpolation Property:** If the design matrix  $X$  has full rank  $k$ , then there exists a global minimizer  $\widehat{\beta}_\theta$  of the CQR distance function such that  $\widehat{\beta}_\theta$  interpolates at least  $k$  data points, i.e. there are  $k$  observations  $\{(y_{i_1}, x_{i_1}), \dots, (y_{i_k}, x_{i_k})\}$  with

$$(IP) \quad y_{i_l} = x_{i_l}'\widehat{\beta}_\theta \text{ for } l = 1, \dots, k \text{ and the rank of } (x_{i_1}, \dots, x_{i_k})' \text{ equals } k.$$

<sup>4</sup>Fitzenberger (1997) treats censoring both from above and below and provides the asymptotic distribution of the CQR estimator. The expression " $\operatorname{sgn}_\theta(\epsilon_i) \cdot \epsilon_i$ " is mostly referred to as "check function"  $\rho_\theta(\epsilon_i)$ . The notation used here has advantages when studying the asymptotic distribution of the estimator.

<sup>5</sup>Womersley (1986, p. 112) and Fitzenberger (1994) provide a more complete characterization of the set of minimizers.

When evaluating the IP, the following three points deserve attention. First, if the CQR distance function exhibits a unique minimizer, it must satisfy IP. Second, the CQR can interpolate a censoring point where an observation is censored. And third, in contrast to standard quantile regressions, Koenker and Bassett (1978), it is not guaranteed that a share of at most  $\theta$   $[(1 - \theta)]$  of the observations lies above [below] the estimated CQR line with an intercept.

The IP is established by analyzing the kinks of the piecewise linear distance function (1) for which the directional derivative proves an important tool. The directional derivative evaluated at  $\beta$  in direction  $w \in \Re^k$  is given by

$$H'_\theta(\beta, w) = \sum_{i=1}^N [I(x'_i\beta < y_{c_i}) \{-sgn_\theta(y_i - x'_i\beta) - I(x'_i\beta = y_i)sgn_\theta(-x'_iw)\} \\ + I(x'_i\beta = y_{c_i}) \{(1 - \theta)I(y_i < y_{c_i}, x'_iw < 0) - \theta I(y_i = y_{c_i}, x'_iw < 0)\}] \cdot x'_iw \quad (2)$$

## 2 Algorithms

Whereas previous simulation studies relied on grid search to determine the CQR estimates, the interpolation property (IP) discussed above suggests an enumeration algorithm to determine the CQR estimate exactly, i.e. an element out of the set of global minimizers. This algorithm, which I denote by IPOL, consists of an enumeration of the set of all  $k$ -tuples of data points with linear independent regressor vectors and the corresponding interpolating regression line. Then the ones minimizing the CQR distance function are in the set of global minimizers. IPOL involves the evaluation of at most  $\binom{N}{k}$   $k$ -tuples. In contrast to grid search, IPOL guarantees to find a global minimum exactly and is typically much faster than grid search. The computational advantage of IPOL relative to grid search increases with the required accuracy of the estimates and the number of the regressors  $k$  and it decreases with the number of observations  $N$ . The algorithms discussed in the following, which are typically much faster than IPOL, will be contrasted in Section 3 with the exact CQR estimates obtained by IPOL.

### 2.1 BRCENS

The algorithm BRCENS is developed in Fitzenberger (1994) as an adaptation of the standard **B**arrodale–**R**oberts–**A**lgorithm (BRA) for standard Quantile Regressions to the **C**ensored Quantile Regression case. A standard quantile regression exhibits a linear programming structure. Barrodale and Roberts (1973) notice that the IP allows for a more efficient,

condensed simplex approach. Only kinks of the distance function need to be considered for which  $k$  (design matrix exhibits full rank) observations are interpolated and for which the rank of the matrix formed by the regressor vectors is equal to  $k$ .

Analogous to BRA, the algorithm BRCENS involves two parts. First, the algorithm starts with all coefficients being zero as the set of nonbasic variables (NB).<sup>6</sup> The algorithm proceeds in  $k$  steps, where in each step that coefficient from NB changes into a direction for which the directional derivative (2) indicates the strongest decline of the objective. This defines a one-dimensional search direction and the coefficient is changed along this direction until the objective starts increasing again. At this point, according to the expression for the directional derivative in (2), there is at least one data point being interpolated. One of these interpolated data points now replaces the coefficient leaving NB. At the end of the first part, the algorithm has reached a situation where IP is satisfied. The second part of BRCENS considers exchanging one of the interpolated observations in NB with a different observation. With all other data points in the NB remaining interpolated, considering one data point defines a one-dimensional search direction. The algorithm keeps moving into the search direction, for which the directional derivative indicates the largest decline of the objective, until the objective cannot be reduced further. At this new point there exists at least one data point which is added to NB, the set of interpolated data points. The algorithm stops when the directional derivatives for all interpolated data points in NB are non-negative, thus guaranteeing that a local minimum has been achieved.<sup>7</sup>

BRCENS does not guarantee convergence to a global minimum of the CQR objective function. The contribution to the directional derivative becomes zero at data points  $i$ , for which the current estimate of the CQR line yields a strictly censored fitted value, i.e.  $x'_i\beta > yc_i$  at the current  $\beta$ . Therefore, the simulation study considers a heuristically modified version of BRCENS, denoted by MBRCENS, which uses different starting values. MBRCENS starts with BRCENS. Based on the set of coefficients obtained, new starting values are used which tend to put the current CQR line out of

---

<sup>6</sup>In the Simplex setup, the set of nonbasic variables comprises those coefficients and residuals (including the corresponding data points) which are currently zero and for which the Simplex tableau provides the linear representation in terms of the variables in the basis. The variables in the basis (coefficients, data points represented by the corresponding residuals) are typically different from zero.

<sup>7</sup>In contrast to the original version of BRCENS, the current version performs the final directional derivative check for all combinations of  $k$  interpolated data points with rank  $k$  as NB. This change improves the reported performance relative to earlier simulation studies, cf. Fitzenberger (1994, 1997).

the censored region. The estimates from the first step are used as starting values for BRCENS in the second step except for the intercept coefficient being shifted to the 25%-quantile of the estimated residuals. Again the estimated coefficients from the second step are modified such that the intercept is now shifted to the 40%-quantile of the estimated residuals and the slope coefficients are multiplied by 0.8 yielding the starting values for a third round with BRCENS. Among the three available estimates, MBRCENS finally takes the one yielding the lowest value of the CQR objective function.

## 2.2 Iterative linear programming algorithm

Buchinsky (1994, p. 412) suggests the following **I**terative **L**inear **P**rogramming **A**lgorithm (ILPA), which has also been applied in Honoré and Powell (1993). ILPA consists of successions of standard quantile regressions. Starting with an initial coefficient estimate  $\hat{\beta}_0$  and a counter  $j = 1$ , the following iterative steps are continued until either convergence is achieved or a maximal number of iterations is reached:

Step 1: For the  $j^{th}$  iteration, determine the set  $M_j$  of observations with  $x'_i \hat{\beta}_j < yc_i$ . If  $j = 1$  or  $M_j \neq M_{j-1}$  then continue with Step 2, otherwise terminate and take  $\widehat{\beta}_\theta = \hat{\beta}_{j-1}$  as the CQR estimate.

Step 2: Calculate  $\hat{\beta}_j$  as the standard quantile regression estimate for the set of observations  $M_j$  by means of the BRA. Set  $j := j + 1$  and repeat Step 1.

Buchinsky states that ILPA is not guaranteed to converge. He motivates ILPA by the following two claims. First, for an optimal solution  $\widehat{\beta}_\theta$  of the CQR problem, the set of observations for which the predicted value lies on or above the censoring point, i.e.  $x'_i \widehat{\beta}_\theta \geq yc_i$ , could have been excluded from the estimation and one would still obtain the same estimate. And second, when convergence is reached, the coefficient estimate represents a local minimum of the CQR distance function.

In Fitzenberger (1994), I provide counter examples showing that there exist both designs for which the CQR estimate interpolates a censoring point and designs for which convergence of ILPA does not result in a local minimum. The main issue is that the contribution of a single observation to the directional derivative (2) changes when the regression line hits the censoring point depending on whether the observation itself is censored and depending on whether the regression line moves towards or away from the censored region.

However, under the assumption that the exact CQR estimate does not interpolate any censored observation, Buchinsky's rationale is true following

from results on standard quantile regressions. Therefore, ILPA has a lot of appeal, since asymptotic consistency of the CQR estimate relies on the fact that population quantiles can be estimated consistently as long as they are uncensored, cf. Powell (1984, 1986).

There exists an alternative version of ILPA, which coincides with ILPA except that in step 1 “ $<$ ” is replaced by “ $\leq$ ” when defining the set  $M_j$ .<sup>8</sup> This allows for the CQR estimates to interpolate censored observations, but again convergence does not guarantee a local minimum, cf. Fitzenberger (1994). Since earlier simulation studies showed that this modification does not lead to an improvement relative to the version of ILPA presented first, it is not considered further in this paper.

### 2.3 NLRQ

The algorithm NLRQ (“Nonlinear Regression Quantile”) developed in Koenker and Park (1996) is a generic interior point algorithm for nonlinear quantile regressions defined by minimizing the distance function  $\sum_{i=1}^N \text{sgn}_\theta[y_i - f_i(x_i, \beta)] \cdot [y_i - f_i(x_i, \beta)]$  with respect to  $\beta$ . The algorithm is built on  $f_i(x_i, \beta)$  being differentiable in  $\beta$  almost everywhere. NLRQ considers successions of linearized quantile regression problems and at each succession the algorithm performs two steps. First, it considers the dual problem to obtain a one-dimensional search direction by interior point methods. Second, the search direction from the dual problem is translated into a search direction in the primal nonlinear problem and a conventional one-dimensional line search is performed. After the two steps, the  $f_i$ ’s and their gradients are updated. The algorithm stops when the new iterate fails to improve the objective function. Koenker and Park provide an S-code, cf. Becker et al. (1988), for NLRQ. For the subsequent simulation study, I have translated their code into Fortran. This makes the timing comparison somewhat unfair for NLRQ since I use grid search for the line search in the second step.

Considering the use of NLRQ for CQR’s, one should note that the  $f_i(\cdot)$ -function is not everywhere differentiable in  $\beta$ . At points satisfying the IP there could be observations for which the regression line interpolates a censoring value involving a kink in  $f_i(\cdot)$ , see the expression for the directional derivative in (2), i.e.  $f_i(\cdot)$  is not differentiable here. Therefore, I also consider a modification of NLRQ (denoted as MNLRQ) in the subsequent simulation study, which takes account of the fact that the directional derivative differs depending on the direction taken when the current CQR

---

<sup>8</sup>In contrast to Buchinsky (1994, p. 412), Buchinsky (1997, section 8.1) refers to this version as the ILPA.

interpolates a censoring point. In such a situation, MNLRQ tries all possible permutations of the contribution of interpolated censoring points to the directional derivative when forming the gradient until a search direction is found along which the CQR objective function can be improved.<sup>9</sup> MNLRQ stops when the CQR objective function cannot be improved further resulting in a local minimum of the CQR optimization problem. It has to be emphasized, that MNLRQ is not constructed to be a serious competitor in terms of computation time. In fact, with a lot of censoring in the data, my current implementation of MNLRQ performs very poorly in this respect, cf. Section 3. The goal here is rather to explore whether NLRQ could be improved by considering more precisely the directional derivative information at censoring points.

### 3 Simulation results

This section analyzes the performance of the algorithms described in Section 2 by means of a simulation study whose design is similar to Fitzenberger (1994, 1997). Table 1 describes the data generating processes (DGP's), (A)–(H). For each scenario, 1000 random samples of size 100 are drawn. The estimation problem is a censored LAD regression with one regressor and an intercept. A sample is dismissed if the exact CQR estimates (determined by the enumeration algorithm IPOL) are not unique.

| DGP | Censoring Values              | True Coefficients                 | Regressor Values               |
|-----|-------------------------------|-----------------------------------|--------------------------------|
| (A) | $yc_i = Const$                | $(\beta_1, \beta_2) = (0, 0)$     | $x_{i,2} \sim N(0, 1)$         |
| (B) | $yc_i = Const$                | $(\beta_1, \beta_2) = (0, 0)$     | $x_{i,2} = -9.9 + 0.2 \cdot i$ |
| (C) | $yc_i = Const + 0.5$          | $(\beta_1, \beta_2) = (0.5, 0.5)$ | $x_{i,2} \sim N(0, 1)$         |
| (D) | $yc_i = Const + 0.5$          | $(\beta_1, \beta_2) = (0.5, 0.5)$ | $x_{i,2} = -9.9 + 0.2 \cdot i$ |
| (E) | $yc_i \sim N(Const, 1)$       | $(\beta_1, \beta_2) = (0, 0)$     | $x_{i,2} \sim N(0, 1)$         |
| (F) | $yc_i \sim N(Const, 1)$       | $(\beta_1, \beta_2) = (0, 0)$     | $x_{i,2} = -9.9 + 0.2 \cdot i$ |
| (G) | $yc_i \sim N(Const + 0.5, 1)$ | $(\beta_1, \beta_2) = (0.5, 0.5)$ | $x_{i,2} \sim N(0, 1)$         |
| (H) | $yc_i \sim N(Const + 0.5, 1)$ | $(\beta_1, \beta_2) = (0.5, 0.5)$ | $x_{i,2} = -9.9 + 0.2 \cdot i$ |

a) *Const* denotes some constant taking various values,  $N(Const, 1)$  denotes the normal distribution with mean *Const* and variance one, and  $I(\cdot)$  denotes the indicator function. The random variables  $\epsilon_i$  are distributed as i.i.d.  $N(0, 1)$  and  $i = 1, \dots, N$

Table 1: Data generating processes (DGP) (A) – (H) used in simulation study for the model  $y_i = \min(yc_i, \beta_1 + \beta_2 \cdot x_{i,2} + \epsilon_i)^a$ .

<sup>9</sup>In cases, when there are too many interpolated censored observations (more than  $4 \times k$ ), each component of the current coefficient vector is perturbed by a very small number. This is done to limit the number of permutations to be explored.

The DGP's differ in four dimensions. First, by whether the coefficients to generate the data are both 0 (A,B,E,F) or both 0.5 (C,D,G,H). Since all algorithms start with both coefficients at 0, it could make a difference whether the starting values are close to the truth. Second, the DGP's differ by whether the censoring points are the same for all observations (A,B,C,D) or differ across observations (E,F,G,H). In the first case, if the CQR line is above the censoring value for a certain regressor value then this is true for all regressor values in a certain neighborhood. In the second case, the CQR line can be censored and uncensored for the same regressor values. This could have an influence on the directional derivative information used locally. Third, the DGP's differ by whether the regressor is a random variable (A,C,E,G) or a fixed sequence of numbers (B,D,F,H). The two scenarios differ by whether all observations exhibit the same a priori distribution. And fourth, the DGP's differ by the degree of censoring depending on  $Const = 1, 0.5, 0$ . Table 2 shows the average share of censored observations depending on the DGP and the choice of  $Const$ .  $Const = 0$  represents a situation where on average 50% of the observations are censored, i.e. the exact CQR ( $\theta = 0.5$ ) typically reaches the censored region.

| DGP | Const = 1.0 | Const = 0.5 | Const = 0.0 |
|-----|-------------|-------------|-------------|
| (A) | 15.9        | 30.9        | 49.7        |
| (B) | 15.9        | 30.9        | 49.8        |
| (C) | 18.7        | 32.9        | 49.9        |
| (D) | 41.0        | 46.0        | 50.9        |
| (E) | 24.1        | 36.4        | 50.1        |
| (F) | 24.1        | 36.3        | 50.0        |
| (G) | 25.4        | 37.1        | 50.0        |
| (H) | 41.0        | 46.0        | 51.0        |

Table 2: Average share of censored observations in random samples for various data generating processes (DGP) – In percent.

**Convergence:** Table 3 presents the absolute frequencies that an algorithm converges. For ILPA, the algorithm is terminated after 20 iterations, if no convergence is achieved. Increasing this number to 100 in some scenarios did not change the results. In such a case, the best coefficient vector during these 20 iterations (in terms of the CQR distance function) is taken as the ILPA estimate.<sup>10</sup> MNLQR and NLRQ are considered not to have converged

<sup>10</sup>In an earlier simulation study, I used the final coefficient estimate after 20 iterations,



| BRCENS, MBRCENS, and NLRQ converged under<br>all DGP's for all 1000 random samples |             |      |             |       |             |       |
|------------------------------------------------------------------------------------|-------------|------|-------------|-------|-------------|-------|
| DGP                                                                                | Const = 1.0 |      | Const = 0.5 |       | Const = 0.0 |       |
|                                                                                    | ILPA        | MLRQ | ILPA        | MNLRQ | ILPA        | MNLRQ |
| (A)                                                                                | 999         | 1000 | 925         | 1000  | 537         | 707   |
| (B)                                                                                | 1000        | 1000 | 962         | 1000  | 508         | 1000  |
| (C)                                                                                | 752         | 1000 | 494         | 1000  | 298         | 989   |
| (D)                                                                                | 662         | 1000 | 656         | 1000  | 642         | 1000  |
| (E)                                                                                | 750         | 1000 | 618         | 1000  | 485         | 1000  |
| (F)                                                                                | 747         | 1000 | 641         | 1000  | 506         | 1000  |
| (G)                                                                                | 727         | 1000 | 621         | 1000  | 526         | 1000  |
| (H)                                                                                | 726         | 1000 | 720         | 1000  | 710         | 1000  |

Table 3: Absolute frequencies among 1000 random samples that algorithms converged.

after 200 iterations. BRCENS, MBRCENS, and NLRQ converge for all scenarios. Convergence is a serious problem for ILPA, especially with a high degree of censoring ( $Const = 0$ ), with bad starting values (C,D) or with random censoring points. When ILPA does not converge, it typically oscillates between two or three coefficient vectors.<sup>11</sup> Additional results (not reported here) indicate that along the iterations ILPA reaches a local minimum as the best coefficient estimate in almost all cases for  $Const = 1, 0.5$  and in at least 76% of all cases for  $Const = 0$ . MNLRQ converges in all cases with low or moderate censoring ( $Const = 1, 0.5$ ) but it exhibits some convergence problems in the presence of a lot of censoring ( $Const = 0$ , DGP (A) and (C)). Overall, lack of convergence is a very serious drawback for the application of ILPA.

**Optimality:** Table 4 is concerned with the frequencies that the global minimum of the CQR distance function is achieved where the latter (the exact CQR estimate) is obtained by the enumeration algorithm IPOL. An algorithm is assumed to have achieved the optimum, if the value of the objective at the solution is within a tolerance of  $10^{-7}$  to the value of the exact

---

cf. Fitzenberger (1994).

<sup>11</sup>If ILPA does not converge, it must oscillate, since a finite sample allows only for a finite number of subsamples which a standard quantile regression can be based upon. Oscillation arises, since an observation, for which the current standard quantile regression implies a censored fitted value, still contributes to the distance function. In the next iteration, this observation is excluded from the sample, which can result in a new estimate for which the fitted value at the aforementioned observation is now uncensored.

CQR estimates. The results are very favorable for BRCENS in comparison. The relative performance of BRCENS is better, when the degree of censoring is higher and when there are more observation specific censoring points. However, for large degrees of censoring, all algorithms (including BRCENS) perform very poorly. Given that BRCENS, MBRCENS, and MNLRQ only guarantee convergence to a local minimum, this poor performance is to be expected, since the CQR distance function is already highly nonconvex with moderate censoring. Unfortunately, a local minimum does not guarantee a solution close to the global minimum, see also the following results on the properties of the coefficient estimates. The modified version MBRCENS yields a substantial improvement compared to BRCENS. ILPA performs better than NLRQ. Again the modified version, MNLRQ improves upon NLRQ and performs better than ILPA in many cases. All algorithms perform quite satisfactorily with moderate degrees of censoring, a common censoring point, and good starting values, DGP (A) and (B) and  $Const = 1$ .

**Properties of Coefficient Estimates:** The quality of the coefficient estimates obtained by various algorithms is an important issue, which has been mostly neglected in my previous simulation studies. Therefore, Tables 5 and 6 also provide results on the root-mean-squared deviation and the 90%-percentile of the absolute deviation of the respective estimates from the exact CQR estimates for moderate degrees of censoring ( $Const = 0.5$ ).<sup>12</sup> According to the root-mean-squared deviations criterion, BRCENS and ILPA exhibit almost the same performance and NLRQ performs noticeably better. This is in contrast to the optimality results presented before. Considering results for  $Const = 0$  (not reported here) reverses the relative performance of BRCENS and NLRQ (BRCENS also outperforms ILPA in this case).<sup>13</sup> The modified algorithm MBRCENS improves upon BRCENS, whereas there is no clear ranking between MNLRQ and NLRQ. MBRCENS performs slightly worse than MNLRQ and NLRQ. Turning to the 90%-percentiles, the numbers are considerably smaller than for the root-mean-squared-deviations, however, the relative performance between the algorithms is only slightly changed (MBRCENS performs better than NLRQ/MNLRQ in a considerable number of cases). Overall, these findings

---

<sup>12</sup>Further results for  $Const = 1, 0.5$  and results on the median and the 75%-percentile are available from the author upon request.

<sup>13</sup>At this point, it remains a topic for further research to investigate why NLRQ proves less outlier sensitive than BRCENS or ILPA for moderate censoring. It seems to pay off that interior point methods do not move too close to the constraint set (in linear programming terminology) in an early stage of the iteration process. Some local minima on the constraint set can actually imply quite extreme coefficient values.

indicate that for all algorithms the distribution of the absolute deviation from the exact CQR estimates is very much skewed to the right. This effect is strongest for BRCENS and ILPA.

| DGP         | ILPA | BRCENS | MBRCENS | NLRQ | MNLRQ |
|-------------|------|--------|---------|------|-------|
| Const = 1.0 |      |        |         |      |       |
| (A)         | 993  | 995    | 995     | 829  | 983   |
| (B)         | 1000 | 1000   | 1000    | 863  | 995   |
| (C)         | 655  | 864    | 908     | 445  | 705   |
| (D)         | 526  | 843    | 938     | 332  | 633   |
| (E)         | 655  | 866    | 904     | 387  | 499   |
| (F)         | 642  | 874    | 932     | 400  | 487   |
| (G)         | 630  | 855    | 913     | 229  | 454   |
| (H)         | 628  | 874    | 958     | 426  | 646   |
| Const = 0.5 |      |        |         |      |       |
| (A)         | 873  | 936    | 942     | 730  | 870   |
| (B)         | 924  | 942    | 958     | 782  | 934   |
| (C)         | 353  | 687    | 745     | 219  | 467   |
| (D)         | 526  | 850    | 932     | 361  | 586   |
| (E)         | 501  | 818    | 861     | 283  | 397   |
| (F)         | 511  | 802    | 876     | 310  | 393   |
| (G)         | 511  | 804    | 884     | 222  | 433   |
| (H)         | 599  | 837    | 936     | 451  | 639   |
| Const = 0.0 |      |        |         |      |       |
| (A)         | 90   | 330    | 379     | 2    | 7     |
| (B)         | 81   | 174    | 408     | 53   | 149   |
| (C)         | 159  | 474    | 596     | 100  | 261   |
| (D)         | 487  | 806    | 911     | 358  | 592   |
| (E)         | 332  | 702    | 774     | 197  | 293   |
| (F)         | 374  | 720    | 845     | 189  | 296   |
| (G)         | 391  | 706    | 809     | 150  | 339   |
| (H)         | 603  | 826    | 942     | 465  | 646   |

Table 4: Absolute frequencies among 1000 random samples that algorithms achieved global optimum of CQR distance function.

**Timing:** Table 7 provides the relative average CPU time requirements of the different algorithms depending on the degree of censoring in the data (*Const*). Timing is of particular importance when bootstrapping the CQR estimates. The results show that BRCENS exhibits the lowest time requirements. ILPA and MBRCENS are next with no clear ranking between the two. In comparison, NLRQ and MNLRQ exhibit a much larger time

| DGP                                   | ILPA  | BRCENS | MBRCENS | NLRQ  | MNLRQ |
|---------------------------------------|-------|--------|---------|-------|-------|
| Const = 0.5 – Estimates for $\beta_1$ |       |        |         |       |       |
| (A)                                   | .565  | .566   | .566    | .475  | .440  |
| (B)                                   | 6.409 | 6.409  | 6.409   | 6.408 | 6.408 |
| (C)                                   | .181  | .179   | .175    | .141  | .146  |
| (D)                                   | .113  | .108   | .075    | .075  | .078  |
| (E)                                   | .045  | .036   | .033    | .029  | .027  |
| (F)                                   | .048  | .043   | .039    | .034  | .038  |
| (G)                                   | .051  | .045   | .037    | .037  | .030  |
| (H)                                   | .121  | .114   | .058    | .073  | .079  |
| Const = 0.5 – Estimates for $\beta_2$ |       |        |         |       |       |
| (A)                                   | .309  | .309   | .309    | .260  | .247  |
| (B)                                   | .662  | .662   | .662    | .662  | .662  |
| (C)                                   | .162  | .156   | .155    | .110  | .122  |
| (D)                                   | .018  | .018   | .012    | .012  | .013  |
| (E)                                   | .051  | .048   | .046    | .037  | .036  |
| (F)                                   | .007  | .007   | .006    | .005  | .006  |
| (G)                                   | .050  | .049   | .041    | .037  | .030  |
| (H)                                   | .019  | .018   | .009    | .013  | .013  |

Table 5: Root-mean-squared deviation of coefficient estimates from exact CQR estimates.

requirement. For a low degree of censoring, BRCENS is about 60 times faster than the exact determination of the CQR estimates by means of IPOL. This advantage is reduced to a factor of 10 when a lot of censoring is present. In comparison, the incremental time requirement for MBRCENS is fairly small, whereas NLRQ and MNLRQ are considerably more expensive. For the case with a lot of censoring, MNLRQ even uses more time than IPOL. However, for fairness sake, it has to be mentioned that my Fortran implementation of NLRQ and MNLRQ is likely to be somewhat inefficient. Given the poor performance of all other algorithms in the presence of a lot of censoring, IPOL appears a viable alternative in such a situation.

**Summary and Recommendations:** Summarizing the simulation results, BRCENS performs quite well in comparison. It outperforms ILPA and NLRQ with respect to the frequencies that the exact CQR estimates are reached. BRCENS is very competitive in terms of the computation times involved and when trying out different starting values its performance can be improved considerably at a fairly small computational cost. All algo-

| DGP                                   | ILPA | BRCENS | MBRCENS | NLRQ | MNLRQ |
|---------------------------------------|------|--------|---------|------|-------|
| Const = 0.5 – Estimates for $\beta_1$ |      |        |         |      |       |
| (A)                                   | .000 | .000   | .000    | .013 | .000  |
| (B)                                   | .000 | .000   | .000    | .011 | .000  |
| (C)                                   | .276 | .256   | .243    | .054 | .030  |
| (D)                                   | .239 | .221   | .000    | .040 | .027  |
| (E)                                   | .108 | .072   | .046    | .049 | .051  |
| (F)                                   | .097 | .084   | .022    | .057 | .051  |
| (G)                                   | .099 | .091   | .030    | .080 | .057  |
| (H)                                   | .267 | .213   | .000    | .050 | .029  |
| Const = 0.5 – Estimates for $\beta_2$ |      |        |         |      |       |
| (A)                                   | .000 | .000   | .000    | .013 | .000  |
| (B)                                   | .000 | .000   | .000    | .002 | .000  |
| (C)                                   | .324 | .267   | .259    | .073 | .045  |
| (D)                                   | .039 | .034   | .000    | .014 | .008  |
| (E)                                   | .112 | .084   | .050    | .051 | .050  |
| (F)                                   | .014 | .017   | .007    | .010 | .010  |
| (G)                                   | .096 | .112   | .032    | .091 | .055  |
| (H)                                   | .043 | .036   | .000    | .012 | .007  |

Table 6: 90%-quantile of absolute deviation of coefficient estimates from exact CQR estimates.

gorithms perform the worse, the higher the degree of censoring. Based on the simulation results reported here and recognizing that BRCENS guarantees convergence to a local optimum, its application can be highly recommended in situations with low or moderate degrees of censoring (relative to the quantile  $\theta$  being estimated). For high degrees of censoring, one might want to determine the exact CQR estimates by means of IPOL. The results for NLRQ show that its performance can be improved when using more precise directional derivative information. However, this could result in considerably higher computation cost (or lack of convergence, when only a small number of iterations is allowed for).

## Acknowledgements

I am grateful to Roger Koenker, Steve Portnoy, and an anonymous referee for helpful comments. All errors are my sole responsibility.

| Const | IPOL  | ILPA | BRCENS | MBRCENS | NLRQ  | MNLRQ  |
|-------|-------|------|--------|---------|-------|--------|
| 1.0   | 66.24 | 1.77 | 1.00   | 1.69    | 24.50 | 21.90  |
| 0.5   | 64.19 | 2.12 | 2.05   | 2.69    | 27.25 | 24.19  |
| 0.0   | 61.96 | 6.86 | 6.41   | 7.27    | 25.55 | 700.87 |

a) The reported numbers are the ratios of average computation times across DGP's (A)–(H) for different degrees of censoring ( $Const = 1, 0.5, 0$ ), cf. table 2, relative to BRCENS,  $Const = 1$ . The time results are obtained with the UNIX 'time' command on an IBM RS 6000 workstation, based on a Fortran implementation of the various algorithms.

Table 7: Average computation times<sup>a</sup>.

## References

- [1] Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press.
- [2] Barrodale, I. and F.D.K. Roberts (1973). An improved algorithm for discrete  $l_1$  linear approximation. *SIAM J. Numer. Anal.* **10** 839–848.
- [3] Becker, R.A., Chambers, J.M., and A.R. Wilks (1988). *The New S Language*. Wadsworth & Brooks/Cole, Pacific Grove.
- [4] Buchinsky, M. (1994). Changes in the U.S. wage structure 1963–1987: application of quantile regression. *Econometrica* **62** 405–458.
- [5] Buchinsky, M. (1997). Recent advances in quantile regression models: a practical guideline for empirical research.. *Journal of Human Resources*. To appear.
- [6] Chamberlain, G. (1994). Quantile regression, censoring, and the structure of wages. In *Advances in Econometrics: Sixth World Congress*, Ed. C. Sims. Econometric Society Monograph, Volume 1.
- [7] Fitzenberger, B. (1994). A note on estimating censored quantile regressions. Center for International Labor Economics, Discussion Paper Nr. 14, University of Konstanz.
- [8] Fitzenberger, B. (1997). A guide to censored quantile regressions. In *Handbook of Statistics*, Volume 15: Robust Inference, Eds. G.S. Madhala, and C.R. Rao. North-Holland. To appear.
- [9] Fitzenberger, B., R. Hujer, T.E. MaCurdy, and R. Schnabel (1995). The dynamic structure of wages in Germany 1976–1984, a cohort analysis. Center for International Labor Economics, Discussion Paper Nr. 22, University of Konstanz.
- [10] Honoré, B.E. and J.L. Powell (1993). Quantile regression under random censoring. Unpublished Manuscript, Princeton University and University of California at Berkeley.
- [11] Horowitz, J. and G. Neumann (1987). Semiparametric estimation of

- employment duration models. *Econometric Reviews* **6** 5–40.
- [12] Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica* **46** 33–50.
  - [13] Koenker, R. and V. D'Orey (1987). Algorithm AS 229. Computing regression quantiles. *Statistical algorithms*, 383–393.
  - [14] Koenker, R. and B.J. Park (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics* **71** 265–283.
  - [15] Nelder, J.A. and R. Mead (1965). A simplex method for function minimization. *Computer Journal* **7** 308–313.
  - [16] Paarsch, H. (1984). A Monte Carlo comparison of estimators for censored regression models. *Journal of Econometrics* **24** 197–213.
  - [17] Powell, J.L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* **25** 303–325.
  - [18] Powell, J.L. (1986). Censored regression quantiles. *Journal of Econometrics* **32** 143–155.
  - [19] Womersley, R.S. (1986). Censored discrete  $l_1$  approximation. *SIAM J. Scient. Statist. Comput.* **7** 105–122.

# On computation of regression quantiles: Making the Laplacian Tortoise faster

Stephen Portnoy

*Department of Statistics, University of Illinois, Champaign, USA*

*Abstract:* In “The Gaussian Hare and the Laplacian Tortoise”, the authors present a two-pronged attack on the computation of  $L_1$  and other regression quantile estimators in linear models for large samples. The first prong involves the application of interior point linear programming methods, specifically designed to treat the absolute error and related regression quantile objective functions. The second prong applies a form of stochastic preprocessing, somewhat reminiscent of the  $O(n)$  algorithms for computing the median of a single sample. These ideas provide computational methods that are in theory faster than least squares as  $n \rightarrow \infty$  (with probability tending to one), and in practice are faster than Spls least squares functions for  $n$  larger than  $10^4$  (and the number of parameters moderate). Here some issues concerning this algorithm are considered, and some improvements are proffered.

*Key words:* Linear models, regression quantiles,  $L_1$ -estimation, computation.

AMS subject classification: 62G05, 62J05, 68A20.

## 1 Introduction

Consider the standard linear model: there are observations  $\{Y_i : i = 1, \dots, n\}$ , satisfying

$$Y_i = x_i' \beta + u_i \quad i = 1, \dots, n \quad (1)$$

where  $x_i$  are vectors in  $\mathbf{R}^p$ ,  $\beta \in \mathbf{R}^p$  is a vector of unknown parameters, and  $\{u_i\}$  form an i.i.d. sequence of errors. Though we consider the model conditionally on  $\{x_i\}$ , we will generally assume that these design vectors are realizations of an independent random process. The traditional approach to



statistical analysis of (1) is to use least squares to estimate the conditional mean of  $Y$  given  $x$ :  $E[Y|x] = x'\beta$ . However, this provides an analysis only of the center of the conditional distribution. To get a more complete picture of the relationship between  $Y$  and  $x$ , Koenker and Bassett (1978) introduced regression quantiles as the solution to the problem: for each  $\tau \in [0, 1]$ , let  $\hat{\beta}(\tau)$  achieve

$$\min_{b \in \mathbf{R}^p} \sum_{i=1}^n \rho_{\tau}(Y_i - x_i' b) \quad (2)$$

where  $\rho_{\tau}(u) = \tau u^+ + (1 - \tau)u^-$ . When  $\tau = .5$ ,  $\hat{\beta}(.5)$  is just the usual  $L_1$  estimator, which corresponds to the conditional median estimator. Under model (1) (and more generally), the line  $y = x'\hat{\beta}(\tau)$  estimates the conditional quantile of  $Y$  given  $x$ . These methods have been applied successfully in a wide variety of examples.

Computation of regression quantile estimators has depended on the recognition that the minimization problem (2) is equivalent to a linear program, *viz.*:

$$\min_{b \in \mathbf{R}^p} \sum_{i=1}^n (\tau u_i + (1 - \tau)v_i) \quad (3)$$

subject to

$$Y_i = x_i' b + u_i - v_i, \quad u_i \geq 0, \quad v_i \geq 0 \quad i = 1, \dots, n$$

This equivalence was first presented in the literature in the mid 1950's. However, it is not unlikely that it was known earlier but dismissed as having little or no computational implication until the power of the simplex algorithm was appreciated. In fact, in the early 19th century Gauss (1809) already recognized that the  $L_1$  estimator could be characterized as having  $p$  zero residuals. As described in Portnoy and Koenker (1997), efficient algorithms based on Danzig's simplex algorithm were developed, and they proved quite effective for sample sizes  $n < 1000$  (or so). However, these algorithms were extremely slow for sample sizes significantly larger than 1000. Thus, Portnoy and Koenker(1997) were led to develop a new computational approach based on two fundamental ideas.

The first idea involved replacing simplex approaches with interior point methods, which originated in the mid 1980's and have been under extremely active development since then. The traditional simplex method is based on the idea that the constraint set for a linear programming problem is a simplex: that is, a convex set defined as the set of convex combinations of a finite number of extreme points (*viz.*, the vertices of the constraint

set). In the regression quantile problem, the vertices are just the  $\beta$ -values defined as having  $p$  zero residuals (which are often called “elemental” solutions). One proceeds iteratively by evaluating the linear objective function at a vertex and finding the adjacent vertex in the direction of steepest descent of the objective function. The solution is found after a finite number of such steps (called “pivots”) when the algorithm reaches a point where no descent direction remains. Unfortunately, in large sample regression quantile problems, the number of vertices is of order  $n^p$ , and this tends to require the algorithm to pass through a very large number of vertices before reaching the solution. Fortunately, under moderate distributional assumptions in model (1), it is possible to show that with probability tending to one, each pivot moves a fraction  $1/n$  towards the solution. Thus, since each pivot takes  $\mathcal{O}(np^2)$  operations, the simplex algorithm might be expected to take  $\mathcal{O}_p(n^2p^2)$  operations (much larger for large  $n$  than the  $\mathcal{O}(np^2)$  operations required for least squares algorithms). In fact, this rate should be reducible to  $\mathcal{O}_p(n^{3/2}p^2)$ , since one can generally obtain a initial estimate within  $\mathcal{O}_p(n^{-1/2})$  of the solution (*e.g.*, the least squares estimate will work as an initial estimate of the median for symmetric distributions). Nonetheless, this is substantially longer than least squares, especially since the constants implicit in the big-O terms are larger for simplex pivoting than for least squares methods.

Modern interior point methods escape this problem by avoiding the boundary of the constraint set. Consider the canonical linear program

$$\min_x \{c'x \mid Ax = b, x \geq 0\} \quad (4)$$

(where  $c$  is a vector,  $A$  is a matrix, and the inequalities are taken coordinate-wise). Associate with this problem the following logarithmic barrier reformulation, which severely penalizes points close to the boundary:

$$\min_x \{B(x, \mu) \mid Ax = b\} \quad (5)$$

where

$$B(x, \mu) = c'x - \mu \sum \log x_k.$$

In effect, (5) replaces the inequality constraints in (4) by the penalty term of the log barrier. Solving (5) with a sequence of parameters  $\mu$  such that  $\mu \rightarrow 0$ , we obtain in the limit a solution to the original problem (4). For each  $\mu$ , problem (5) can be solved relatively effectively by iterative Newton or quasi-Newton methods: at each trial solution, one approximates the problem locally by a quadratic minimization problem and moves as far as possible towards the solution of the approximating problem without leaving

the constraint set (that is, remaining in the interior). The historical context of these methods and the details of our use of this approach for regression quantiles is presented in Portnoy and Koenker (1997). It appears that such methods can be as reliable as simplex methods, and are substantially faster for large  $n$ .

Nonetheless, the interior point algorithms were still much slower than least squares for very large  $n$ . In fact, the best rates for complexity available from the interior point literature are of the order  $\mathcal{O}_p(n^{5/4}p^3)$  computer operations for random problems, though it has been conjectured that the  $n^{5/4}$  factor can be reduced to  $n \log n$ . The second approach to accelerating the computation of regression quantiles is based on a stochastic preprocessing step that begins with a much smaller random subset of the data. To describe this approach, consider the  $L_1$  problem:

$$\min_b \sum_{i=1}^n |Y_i - x'_i b|. \quad (6)$$

Suppose for the moment that we “knew” that a certain subset  $J_H$  of the observations  $N = \{1, \dots, n\}$  would fall above the optimal median plane and another subset  $J_L$  would fall below. Then, since knowing the sign of a residuals permits the replacement of the absolute value by the appropriate sign,

$$\sum_{i=1}^n |Y_i - x'_i b| = \sum_{i \in N \setminus J_L \cup J_H} |Y_i - x'_i b| - \sum_{i \in J_L} (Y_i - x'_i b) + \sum_{i \in J_H} (Y_i - x'_i b).$$

It follows that

$$\sum_{i=1}^n |Y_i - x'_i b| = \sum_{i \in N \setminus J_L \cup J_H} |Y_i - x'_i b| + |Y_L - x'_L b| + |Y_H - x'_H b| \quad (7)$$

where  $x_K = \sum_{i \in J_K} x_i$ , and  $Y_K = \sum_{i \in J_K} Y_i$  for  $K \in \{H, L\}$ . We will refer to these combined “pseudo-observations” as “globs” in what follows. It is not hard to show that minimizing (7), under our provisional hypothesis on the signs of the residuals, yields exactly the same solution as (6), but the revision has reduced effective sample size by  $\#\{J_L \cup J_H\} - 2$  (essentially by the number of observations in the globs).

To find  $J_L$  and  $J_H$ , consider computing a preliminary estimate  $\hat{\beta}$  based on a subsample of  $m$  observations. Compute a simultaneous confidence band for  $x'_i \beta$  based on this estimate for each  $i \in N$ . Under plausible sampling assumptions the length of each interval is proportional to  $1/\sqrt{m}$ , so if  $M$  denotes the number of  $Y_i$  falling inside the band,  $M = \mathcal{O}_p(n/\sqrt{m})$ . Take  $J_L, J_H$  to be composed of the indices of the observations falling

outside the band. So we may now create the “globbed” observations  $(Y_K, x_K), K \in \{L, H\}$  and reestimate based on  $M + 2$  observations. Finally, we must check to verify that, in fact, all the observations in  $J_H, J_L$  have the predicted residual signs. If so, we are done; if not, we must repeat the process. If the coverage probability of the bands is  $P$ , presumably near 1, then the expected number of repetitions of this process is the expectation of a geometric random variable,  $Z$ , with expectation  $P^{-1}$ . We will call each repetition a cycle. As described in Portnoy and Koenker (1997), it is possible to show that the optimal choice for the initial subsample size is  $m = \mathcal{O}(n^{2/3})$ , and that the resulting size of the globbed sample is then  $M = \mathcal{O}_p(n^{2/3})$ . Under modest distributional assumptions, this provides an algorithm with complexity  $\mathcal{O}_p(n^{2/3}(\log n)^2 p^3) + \mathcal{O}(np)$ , where the last term comes from the computation of the globs and the checking of residuals. For  $n$  large and  $p$  moderate, this is strictly better than the rate  $\mathcal{O}(np^2)$  for least squares; and in fact the algorithm in Portnoy and Koenker (1997) is essentially as fast as Splus least squares algorithms for  $n < 10^6$  and  $p$  moderate; and it is undoubtedly strictly faster for  $n$  much larger than this range.

The algorithm presented in Portnoy and Koenker (1997) can be described formally, but briefly, as follows:

```

 $k \leftarrow 0$ 
 $l \leftarrow 0$ 
 $m \leftarrow \lceil 2n^{2/3} \rceil$ 
while( $k$  is small){
   $k = k + 1$ 
  solve for initial rq using first  $m$  observations
  compute confidence interval for this solution
  reorder globbed sample as first  $M$  observations
  while( $l$  is small){
     $l = l + 1$ 
    solve for new rq using the globbed sample
    check residual signs of globbed observations
    if no bad signs: return optimal solution
    if only few bad: adjust globs, reorder sample, update  $M$ , continue
    if too many bad: increase  $m$  and break to outer loop
  }
}
```

Here, “rq” is the regression quantile problem, and an interior point

method is used to “solve” this problem. The remainder of the paper describes several approaches to improving this algorithm and to obtaining a better understanding of its performance.

## 2 Some simple improvements

Three relatively simple and straightforward improvements in the algorithm above have been made and tested. The first concerns the choice of the initial random subsample. The algorithm above assumes that the sample comes already randomized. For samples that are not simulated, this requires an initial random permutation of all the data – a rather time-consuming task if  $n$  is very large. Note that the above algorithm does not permit simply taking an initial random subsample of size  $m$ , since  $m$  increases with each cycle in order to ensure termination of the iterations. One could take a random subsample of size  $m(k)$  (where  $k$  is the cycle number) at the beginning of each cycle, but this is still rather time-consuming. A faster approach is to form the “random” subsample by taking a random observation from each consecutive  $n/m$  observations. Although this is slightly different from taking a fully random permutation, it appears to be sufficiently random in all cases checked so far. It is extremely quick, requiring only  $m$  random uniforms and no sorting. It has the added advantage for real data of avoiding any large gaps in the sampling (as can occur with fully random permutations). The current implementation uses a simple multiplicative congruential generator. It appears to provide a modest improvement in timings at the cost only of keeping track of an additional random seed.

The second relatively simple improvement involves choice of the simultaneous confidence bands used to determine the residual signs for the first subsample. The algorithm above was originally programmed using the traditional Scheffé bands of the form

$$x_i' \hat{\beta} \pm \left( c x_i' (X' X)^{-1} x_i \right)^{1/2} / \hat{\sigma}_\tau$$

where  $c$  is a constant (from F-tables) and  $\hat{\sigma}_\tau^2$  is an estimate of  $\tau(1 - \tau)/f^2(F^{-1}(\tau))$ . Unfortunately, these bands require  $np^2$  operations, a value that can make the algorithm very time-consuming and, in fact, does not even attain the complexity rate claimed above. To get a faster approach, note that it is preferable to choose the constant  $c$  to optimize the speed of the algorithm rather than to attain a given coverage probability. Thus, different conservative alternatives might prove better. One method that seems to work well is based on the inequality,

$$|x'_i \hat{\beta}| \leq \max_j \left\{ |\hat{\beta}_j| / s_j \right\} \times \sum_{j=1}^p |x_{ij}| s_j,$$

where  $s_j$  is  $\hat{\sigma}$  times the diagonal element of the  $(X'X)^{-1}$  matrix, and  $\hat{\sigma}$  computed as for the Scheffé intervals. This approach provides conservative (though not “exact”) confidence bands with width  $c_q \sum_{j=1}^p |x_{ij}| s_j$ . Note that this requires only  $\mathcal{O}(np)$  operations; thus providing the rate required for the result of Section 1. Choice of the constant,  $c_q$ , is somewhat problematic, but some experimentation with simulated data showed that  $c_q$  could be taken conservatively to be approximately one, and that the algorithm was remarkably independent of the precise value of  $c_q$ . Although our initial experience with this approach is extremely promising, the next improvement discussed below permits the selection of  $c_q$  to be replaced by a simpler selection of an alternative parameter.

A third improvement concerns the size of the globbed sample. If the constant,  $c$ , in the simultaneous confidence intervals is fixed, the size of the globbed sample,  $M$ , is random. Since it is optimal to have  $M$  approximately equal to  $m$ , where  $m$  is the initial sample size, we have tried to choose the  $c$  so that  $M$  is near  $m$  with high probability (this is possible under model (1) as  $n \rightarrow \infty$ ). In practice,  $M$  varies significantly, and this leads to difficulties. If  $M$  is too small, the final estimates are likely to be wrong, thus requiring extra cycles. If  $M$  is too large, the interior point algorithm takes much more time than it should. Since in the limit,  $M$  will be very near its mean, which is a constant times  $m$ , it seems reasonable to fix the sample size  $M \equiv am$ , where  $a$  is a constant near 1. The globbed sample consists of the  $M = am$  observations with the smallest values of  $|r_i|/z_i$ , where  $r_i$  is the residual and  $z_i$  is the constant in the confidence intervals depending only on the design matrix. That is,  $z_i = (x'_i(X'X)^{-1}x_i)^{1/2}$  for Scheffé intervals, and  $z_i = \sum_{j=1}^p |x_{ij}| s_j$  for the conservative intervals described above. This is asymptotically equivalent to fixing  $c$ , but always provides an appropriate globbed sample size. In a small scale simulation study, it appeared that  $a = .8$  was a good choice over a variety of data set sizes and distribution assumptions. The lack of randomness in  $M$  provides a much more reliable and faster algorithm, whose performance varies substantially less from trial to trial. The result that  $a < 1$  probably arises from the fact the the globbed sample is not like a random sample (as discussed in Section 4). Thus the interior point algorithm takes somewhat more time on a globbed sample than on a random subsample of the same size (perhaps 20 to 40 per cent more time, depending of the specific problem).

### 3 On the interior point algorithm

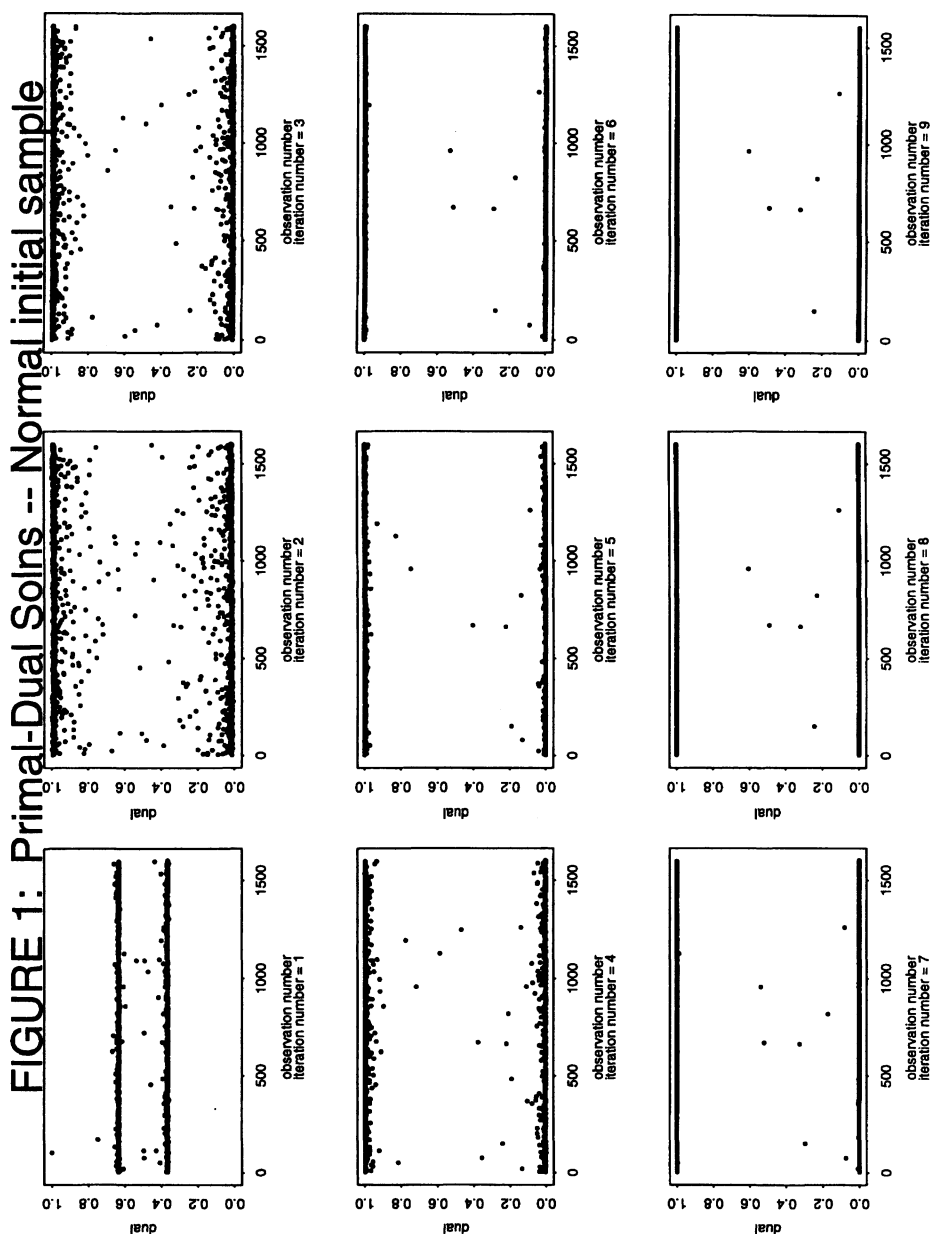
For  $n < 10^6$ , the algorithm of Section 1 spends almost all of its time in the interior point optimization steps for the random and globbed subsamples. However, the full advantage of the preprocessing step isn't realized until the operations of complexity  $\mathcal{O}(np)$  dominate (that is, until most of the time is spent in creating the globs and in checking residuals). Thus, improvements in the interior point algorithm should provide the most important sources of faster performance (for  $n < 10^6$ ). Furthermore, the use of nonlinear optimization methods requires the specification of a number of performance parameters. These include: (i) choice of starting values for the variables over which optimization is carried out, (ii) choice of direction at each step (for example, use of gradient direction for steepest descent, or use of the Newton direction to solve the approximating quadratic, or some combination of these), (iii) specification of the distance to move along the chosen descent direction, (iv) selection of a method for updating the penalty coefficient  $\mu$  in equation (5), and (v) selection of stopping criteria. Specification of these parameters provides ample room for fine-tuning the algorithm to provide better performance.

Interior point methods for  $L_1$  problems generally replace the linear program in equation (3) by a related problem, called the dual problem

$$\max\{Y'a \mid X'a = \frac{1}{2}X'e, \ a \in [0, 1]^n\}. \quad (8)$$

where  $Y$  is the vector of response observations and  $X$  is the design matrix. Here the coordinates of  $a$ ,  $\{a_i\}$ , correspond to the signs of the residuals,  $r_i \equiv Y_i - x_i'b^*$ , at the final solution,  $a^*$ , where  $b^*$  is the optimal solution to equation (3). Precisely, at a solution,  $a_i = 1$  if  $r_i > 0$ ,  $a_i = 0$  if  $r_i < 0$ , and  $a_i$  is strictly between 0 and 1 if  $r_i = 0$ . In the regression quantile setting, the values  $a_i(\tau)$  ( $0 \leq \tau \leq 1$ ) are exact analogues of the rank functions of Hájek and Šidák (1967): see Gutenbrunner and Jurečková (1992). Thus, the solution  $b^*$  to the primal problem (3) can be determined from the solution to (8); and, in fact, the objective functions are the same at the solutions. The “primal-dual” algorithm of Portnoy and Koenker (1997) proceeds by establishing first order conditions depending on both  $a$  and  $b$ , and solving for both simultaneously. That is, in each iteration, both  $a$  and  $b$  are moved in a Newton direction toward the solution. At each iteration, the difference between the primal and dual objective functions is positive until the solution is reached. At the solution, this difference, called the “duality gap” becomes zero, and this provides a reliable test for convergence of the algorithm. Although each iteration of the primal-dual algorithm is somewhat more complicated, the number of iterations tends

to be somewhat smaller than other approaches, and the algorithm appears to be quite robust to idiosyncrasies in the data.



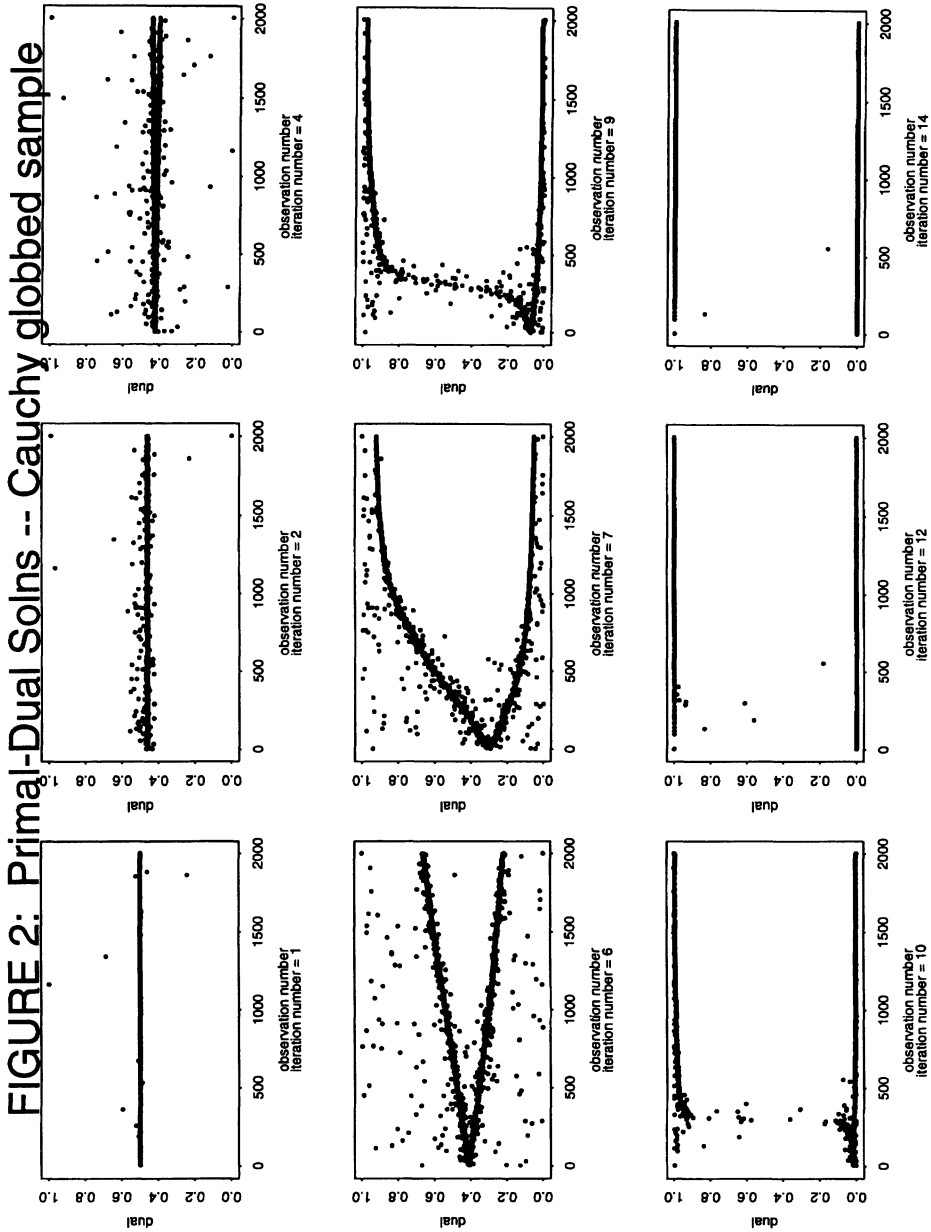
The algorithm in Portnoy and Koenker (1997) chooses the initial starting values as follows:  $b_0$  is chosen to be the least squares estimator, and  $a_0$  is chosen to be a constant vector with all  $n$  elements equal to .5 (halfway



between the bounds 0 and 1). In an effort to see how this algorithm works, the values of the dual variables,  $a_i$ , were plotted against  $i$  at each iteration. Plots were made for several random data sets with sample size 1000 or 2000, values near typical initial sample sizes for problems with  $n$  between  $10^4$  and  $10^5$ . One picture of successive iterations on normal data with  $p = 6$  appears in Figure 1. The first plot gives the  $a$ -values after one iteration. Note that the values have separated from the initial constant .5 and have begun to move toward 0 and 1. Remarkably, the vast majority of the values move exactly the same fraction of the distance to the extreme limits, 0 and 1. The values quickly approach the limiting values, except for the observations with zero residuals at the solution. The last iteration represents a very minor fine tuning of the next to last one, which already identified the zero residuals. This suggests the possibility of stopping a bit earlier, but in practice early stopping provides only a modest improvement in timings (at the cost of potentially less reliable performance).

Similar plots were made when the primal-dual algorithm was applied to the globbed sample. Here the  $a_i$ -values were plotted against  $|r_i|/z_i$ , where  $r_i$  are the residuals and the  $z_i$ -values are defined in the third improvement discussed in Section 2. That is, the  $a_i$  are plotted against the order at which observations enter the globbed sample. A picture for Cauchy data with  $p = 3$  is given in Figure 2. At first glance, the results appear even more remarkable. With the residual ordering taken into account, it is clear that the  $a_i$ -values for smaller residuals tend to their limits much more slowly than those for larger residuals (among the globbed sample, for which all but the last two observations – the globs – have small residuals). Again, the  $a_i$ -values tend to fall along lines, but the lines are not symmetric about .5, and the become highly curved after iteration 6.

It is possible that the use of the least squares estimator as the initial starting value for applying the interior point algorithm to the globbed sample might be responsible for some of the oddities in the plots. An alternative that should be somewhat better and might be somewhat faster is to use the  $\hat{\beta}$  from the solution to the initial random subsample as the starting value for the globbed sample. It turns out that this does not affect the plots of the  $a_i$ -values for the specific example plotted in Figure 2. However, in various simulation experiments, use of the initial  $\hat{\beta}$  as the starting value for applying the interior point algorithm to the globbed samples appeared to provide a modest but definite improvement. These figures give some tantalizing hints as to why this is so, but significantly better understanding of the performance of interior point methods should provide even more substantial improvements.



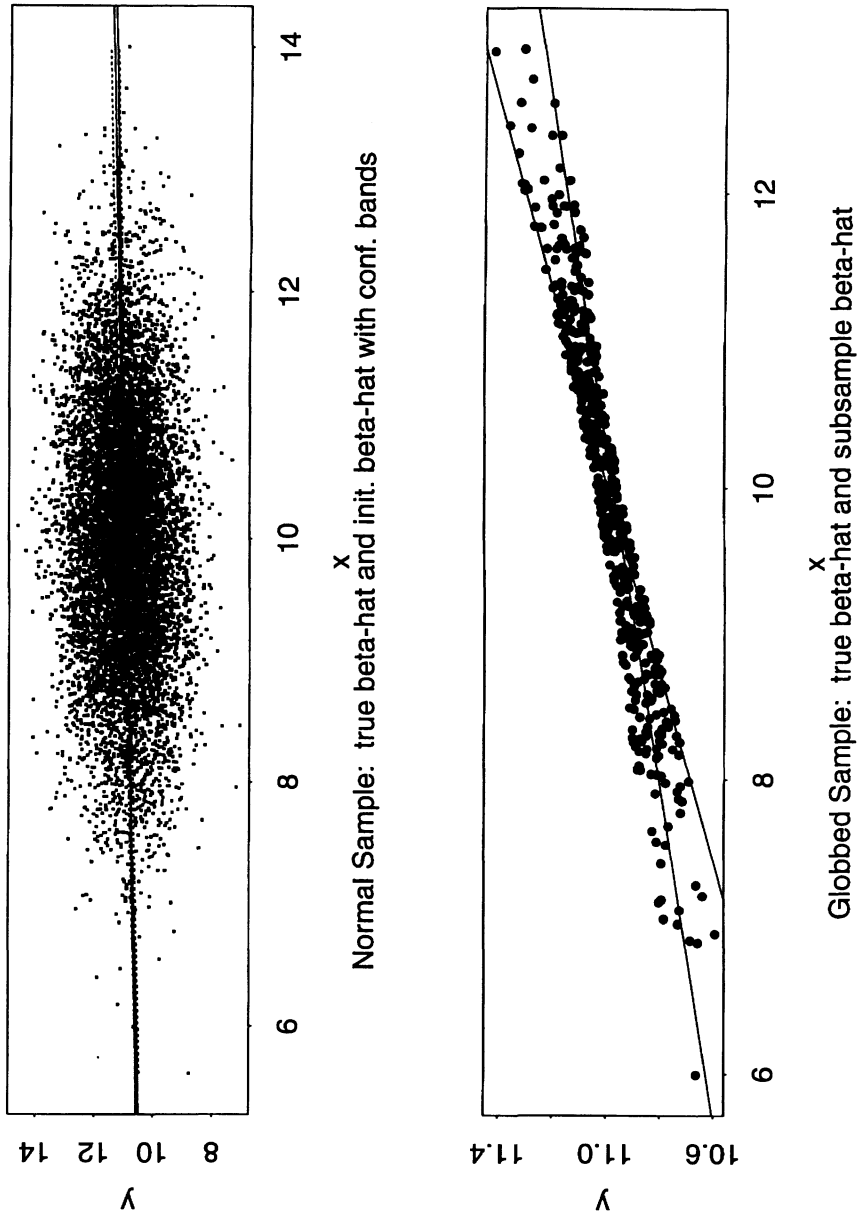
#### 4 On applying the algorithm recursively

An obvious source of improved performance for very large samples would clearly be to apply the algorithm recursively. Since the algorithm gives the fastest computation of regression quantiles, the use of interior point meth-

ods should be replaced by the full algorithm using preprocessing whenever the subsample sizes are sufficiently large to make this replacement noticeably faster. Generally, this would occur when  $n$  is somewhat larger than  $10^5$ . Unfortunately, there is a serious problem with this replacement for solving the globbed problem. The stochastic preprocessing step assumes that the sample is a random one. The globbed sample is far from random – it is chosen to consist of the smaller residuals from the initial sample plus the two globs. Thus, the preprocessing step cannot be expected to provide any real reduction in sample size for the globbed sample. Figure 3 should make this clear. The first graph in Figure 3 plots a Normal sample of size 10,000 together with the the whole-sample  $L_1$  line, the initial subsample  $L_1$  line and the confidence bands (based on the subsample). The solution to the initial subsample should differ from the correct sample regression quantile by an error of order  $\mathcal{O}_p(m^{-1/2})$  (where  $m$  is the initial subsample size). This error is of the same order as the width of the confidence bands. Therefore, as the first graph shows, the lines and confidence bands are extremely close together on the scale of the data. The second plot of Figure 3 shows just the globbed sample together with the correct  $L_1$  line and an  $L_1$  line based on a random subsample of the globbed sample. Clearly, any confidence band about this subsample line that contains the correct line must contain all but a relatively modest fraction of the globbed sample. That is, since the residuals should be roughly uniformly distributed between the bands (assuming the random errors have a smooth density near the desired quantile), it is clear that an appreciable fraction of them must have their residuals from a subsample estimate differ in sign from those of the correct estimate. Thus, for the globbed sample, it would be impossible to replace the problem by one with a sample size smaller than a constant fraction of the globbed sample size. This contrasts markedly with the reduction from  $n$  to  $n^{2/3}$  that preprocessing affords for random samples.

It is possible to use the preprocessed algorithm for the initial random sample. A few simulations with  $n > 10^6$  and  $p \leq 4$  were tried with this modification, and a modest improvement (about 20%) was obtained. Unfortunately, computer space limitations precluded more extensive testing, which remains to be done.

FIGURE 3: Picture of Required Subsample Size



## References

- [1] Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium ... ; Werke*. Gotha (1871), Second Book, Third part, Section 186.
- [2] Gutenbrunner, C. and Jurečková, J. (1992). Regression rank scores and

regression quantiles. *Ann. Statist.* **20** 305-330.

- [3] Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. Prague: Academia.
- [4] Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46** 33-50.
- [5] Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: computability of squared error and absolute-error estimators. *Statistical Science*. To appear.

# Recent developments in PROGRESS

Peter J. Rousseeuw and Mia Hubert

*University of Antwerp, Belgium*

*Abstract:* The least median of squares (LMS) regression method is highly robust to outliers in the data. It can be computed by means of PROGRESS (from Program for RObust reGRESSion). After ten years we have developed a new version of PROGRESS, which also computes the least trimmed squares (LTS) method. We will discuss the various new features of PROGRESS, with emphasis on the algorithmic aspects.

*Key words:* Algorithm, breakdown value, least median of squares, least trimmed squares, robust regression.

AMS subject classification: 62F35, 62J05.

## 1 Introduction

At the time when the least median of squares (LMS) regression method was introduced (Rousseeuw, 1984), a program was needed to compute it in practice. The first algorithm described in that paper was just for computing the LMS line in simple regression, based on scanning over possible slopes while adjusting the intercept each time.

However, it was clear from the start that an algorithm for LMS multiple regression was required. The first version of PROGRESS (from Program for RObust reGRESSion) was implemented in 1983. The 1984 paper already contained an example analyzed with PROGRESS and listed the program's computation times on a CDC 750, one of the fastest mainframes of that day but outperformed by today's PC's. During the next years, when people began requesting the program, it was made more user-friendly with interactive input and self-explanatory output. The use of the program was explained in detail in (Rousseeuw and Leroy, 1987). Because that book contained many sample outputs we refrained from making any substantial modifications to PROGRESS, which remained essentially unchanged from

1985 until 1995.

During that decade there were quite a few suggestions for modifications and extensions. For instance, several people asked for the inclusion of the least trimmed squares (LTS) method which had been proposed together with LMS in (Rousseeuw, 1984) but which was not built in from the start because it needed (a little) more computation time, which became less relevant with the increasing speed of hardware. Another idea was to improve the accuracy by carrying out intercept adjustments more often, and we also wanted to allow the user to replace the ‘median’ in LMS by another quantile. Therefore we finally gave up on the principle of keeping the outputs identical to those in the 1987 book, and created the modernized 1996 version of PROGRESS described in the present paper.

First of all, PROGRESS now allows the user to choose between two robust estimators: the least quantile of squares (LQS) method which generalizes LMS, and the least trimmed squares (LTS) method. By definition, these methods depend on a quantile  $h/n$ . In order to help the user make an appropriate choice of  $h$ , the program provides a range of  $h$ -values for which LQS and LTS have a breakdown value between 25% and 50%. (This means that the method can resist that many contaminated observations.) Section 2 describes the LQS and LTS, and Section 3 obtains their breakdown value which depends on  $h$ .

Section 4 provides an outline of the algorithm used for the LQS and LTS. Since their objective functions are difficult to minimize exactly, PROGRESS performs an approximate resampling algorithm. Whereas the 1985 version adjusted the intercept only once at the end, the intercept is now adjusted in each step. This yields a lower objective function value, and in simple regression we even find the exact minimum. The program now allows to search over all subsets, as well as over a user-defined number of random subsets.

In Section 5 we define a new version of the robust coefficient of determination ( $R^2$ ) to make sure that it always takes on values in the interval  $[0,1]$ . Finally, Section 6 discusses the robust diagnostics which the program provides to identify outliers and leverage points. Section 7 explains how the program can be obtained.

## 2 The estimators LQS and LTS

We consider the linear multiple regression model

$$y_i = x_{i1}\theta_1 + x_{i2}\theta_2 + \dots + x_{ip}\theta_p + \sigma e_i = \mathbf{x}_i^t \boldsymbol{\theta} + \sigma e_i \quad (1)$$

for  $i = 1, \dots, n$ . The  $p$ -dimensional vectors  $\mathbf{x}_i$  contain the explanatory

variables,  $y_i$  is the response and  $\sigma e_i$  is the error term. The data set thus consists of  $n$  observations and will be denoted by  $Z = (X, \mathbf{y})$ . For a given parameter estimate  $\hat{\theta}$  we denote the residuals as  $r_i(\hat{\theta}) = y_i - \mathbf{x}_i^t \hat{\theta}$ . In a regression model with intercept, the observations satisfy  $x_{ip} \equiv 1$ .

A robust regression method tries to estimate the regression parameter vector  $\theta$  in such a way that it fits the bulk of the data even when there are outliers.

The new version of PROGRESS provides two such robust regression methods: the least quantile of squares (LQS) and the least trimmed squares (LTS) estimator. Whereas classical least squares (LS) minimizes the sum of the squared residuals, LQS and LTS minimize a certain quantile, resp. a trimmed sum, of the squared residuals. Their exact definition is given below. (For any numbers  $u_1, \dots, u_n$  the notation  $u_{i:n}$  stands for the  $i$ -th order statistic.)

**Definition 1** Let  $Z = (X, \mathbf{y})$  be a data set of  $n$  observations in  $\mathbb{R}^{p+1}$ . Then for all  $p \leq h \leq n$ , the least quantile of squares (LQS) estimate  $\hat{\theta}_{LQS}(Z)$  and the least trimmed squares (LTS) estimate  $\hat{\theta}_{LTS}(Z)$  are defined by

$$\hat{\theta}_{LQS}(Z) = \underset{\theta}{\operatorname{argmin}}(r^2(\theta))_{h:n} = \underset{\theta}{\operatorname{argmin}}|r(\theta)|_{h:n} \quad (2)$$

and

$$\hat{\theta}_{LTS}(Z) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^h (r^2(\theta))_{i:n}. \quad (3)$$

It is easy to see that LQS generalizes the LMS method which minimizes the median of the squared residuals. Indeed, for  $n$  odd and  $h = [n/2] + 1$  the LQS becomes the LMS.

With the parameter estimates  $\hat{\theta}_{LQS}(Z)$  and  $\hat{\theta}_{LTS}(Z)$  we can associate estimators of the error scale  $\sigma$ :

$$s_{LQS}(Z) = s_{LQS}(X, \mathbf{y}) = c_{h,n} |r(\hat{\theta}_{LQS}(Z))|_{h:n} \quad (4)$$

and

$$s_{LTS}(Z) = s_{LTS}(X, \mathbf{y}) = d_{h,n} \sqrt{\frac{1}{h} \sum_{i=1}^h (r^2(\hat{\theta}_{LTS}(Z)))_{i:n}}. \quad (5)$$

The constants  $c_{h,n}$  and  $d_{h,n}$  are chosen to make the scale estimators consistent at the gaussian model, which gives

$$\begin{aligned} c_{h,n} &= 1/\Phi^{-1}\left(\frac{h+n}{2n}\right) \\ d_{h,n} &= 1/\sqrt{1 - \frac{2n}{hc_{h,n}}\phi(1/c_{h,n})}. \end{aligned}$$



Moreover,  $s_{LQS}$  is multiplied by a finite-sample correction factor. For  $h = [n/2] + 1$  this factor equals  $1 + \frac{5}{n-p}$ .

More efficient scale estimates, based on the preliminary ones, are then given by

$$\hat{\sigma} = \sqrt{\frac{\sum_i w_i r_i^2}{\sum_i w_i - p}}, \quad (6)$$

where

$$w_i = \begin{cases} 0 & \text{if } |r_i/s_{LQ(T)S}| > 2.5 \\ 1 & \text{otherwise.} \end{cases}$$

Here the notation  $s_{LQ(T)S}$  stands for  $s_{LQS}$  or  $s_{LTS}$ , whichever is used.

### 3 Breakdown value and choice of $h$

In the next theorem we derive the breakdown value of LQS and LTS, which says how many of the  $n$  observations need to be replaced before the estimate is carried away. The finite-sample breakdown value (Donoho and Huber, 1983) of any regression estimator  $T(Z) = T(X, \mathbf{y})$  is given by

$$\varepsilon_n^* = \varepsilon_n^*(T, Z) = \min \left\{ \frac{m}{n}; \sup_{Z'} \|T(Z')\| = \infty \right\}$$

where  $Z' = (X', \mathbf{y}')$  ranges over all data sets obtained by replacing any  $m$  observations of  $Z = (X, \mathbf{y})$  by arbitrary points. We will assume that the original  $X$  is in *general position*. This means that no  $p$  of the  $\mathbf{x}_i$  lie on a  $p - 1$  dimensional plane through the origin. For simple regression with intercept ( $p = 2$ ) this says that no two  $x_i$  coincide. For simple regression without intercept ( $p = 1$ ) it says that none of the  $x_i$  are zero.

**Theorem 1** *If the  $\mathbf{x}_i$  are in general position, then the finite-sample breakdown value of the LQS and the LTS is*

$$\varepsilon_n^* = \begin{cases} (h - p + 1)/n & \text{if } p \leq h < \lceil \frac{n+p+1}{2} \rceil \\ (n - h + 1)/n & \text{if } \lceil \frac{n+p+1}{2} \rceil \leq h \leq n. \end{cases}$$

The proof is given in the Appendix.

**Corollary 1** *If the  $\mathbf{x}_i$  are in general position, then the maximal finite-sample breakdown value of the LQS and the LTS equals*

$$\max_h \varepsilon_n^* = \frac{[(n-p)/2] + 1}{n}$$

and is achieved for

$$[(n+p)/2] \leq h \leq [(n+p+1)/2].$$

When  $n+p$  is even we have  $[(n+p)/2] = [(n+p+1)/2]$  hence the optimal  $h$  is unique. When  $n+p$  is odd, it turns out that choosing  $h = [(n+p+1)/2]$  gives the better finite-sample efficiency. Therefore, we will always define the optimal  $h$  as

$$h_{opt} = [(n+p+1)/2].$$

This is also the default value of  $h$  in PROGRESS. If the user prefers to use another quantile, the program displays a range of  $h$ -values for which a breakdown value of at least 25% is attained. The lowest  $h$ -value allowed in the program is

$$h_{min} = [n/2] + 1.$$

(This is because for each  $\tilde{h} < h_{min}$  there exists some  $h \geq h_{min}$  with the same breakdown value and a higher finite-sample efficiency.)

**Remark 1** *If  $p = 1$  and  $x_{ip} = 1$  for all observations, the regression model reduces to the univariate model  $y_i = \mu + \sigma e_i$ . In that case Theorem 1 is still valid, whereas the LQS and LTS become much easier to compute. In the univariate setting, a fast algorithm is available to compute the exact LQS and LTS estimates of the location parameter  $\mu$  and the scale parameter  $\sigma$ .*

|                                   |
|-----------------------------------|
| 1. input the data and all options |
| 2. treatment of missing data      |
| 3. standardize the data           |
| 4. LS analysis                    |
| 5. compute LQS or LTS             |
| 6. RLS analysis                   |

Table 1: Overview of the program PROGRESS.

## 4 Outline of the PROGRESS algorithm

PROGRESS not only computes the LQS and LTS. First, the LS estimates and inferences about the regression parameters are obtained. And after the LQS or LTS is found, a reweighted least squares (RLS) is carried out with weights based on LQS or LTS. Table 1 gives a schematic overview of the complete program. Since the essential algorithmic changes have been made in step 5, we will focus on that part here. We refer to (Rousseeuw and Leroy, 1987) for all details about the treatment of missing data, the standardization procedure, and the LS and RLS estimates.

|   | <b>ACTION</b>                                                                         | <b>RESULT</b>                                                           |
|---|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------|
| 1 | draw a (random) subset of $p$ observations                                            | $\{z_{i_1}, \dots, z_{i_p}\}$                                           |
| 2 | compute hyperplane through these $p$ observations                                     | $\theta = (\theta_1, \dots, \theta_{p-1}, \theta_p)$                    |
| 3 | if regression with intercept<br>$\Rightarrow$ adjust intercept                        | $\tilde{\theta} = (\theta_1, \dots, \theta_{p-1}, \theta'_p)$           |
| 4 | evaluate the objective function at this estimate                                      | $ r(\tilde{\theta}) _{h:n}$ or $\sum_{i=1}^h r^2(\tilde{\theta})_{i:n}$ |
| 5 | repeat steps 1 until 4, and keep the<br>estimate with lowest objective function value | $\hat{\theta}_{LQS}$ or $\hat{\theta}_{LTS}$                            |

Table 2: Summary of the algorithm for LQS and LTS.

In general the objective functions of LQS and LTS are difficult to minimize exactly since they have several local minima. For this reason PROGRESS uses an approximate resampling algorithm (which does yield the exact solution in simple regression). Table 2 summarizes the main steps of this algorithm.

| $n$          | <b>mechanism</b>      | <b>number of <math>p</math>-subsets used</b> |
|--------------|-----------------------|----------------------------------------------|
| small        | all subsets           | $C_n^p$                                      |
| intermediate | all subsets<br>random | $C_n^p$<br>default (Table 4) or user-defined |
| large        | random                | default or user-defined                      |

Table 3: Subsampling mechanism in PROGRESS.

We will describe the first three steps more extensively.

### 1. draw a (random) subset of $p$ observations

The drawing mechanism now implemented in PROGRESS is displayed in Table 3. According to the sample size  $n$  and the number of variables  $p$ , PROGRESS checks whether or not it is feasible to draw all subsets of  $p$  observations out of  $n$ .

For **small** values of  $n$  (see Table 4) the program automatically generates all possible subsets of  $p$  observations, of which there are  $C_n^p = \binom{n}{p}$ . For each of these  $p$ -subsets, steps 1 to 4 of Table 2 are carried out.

If  $n$  is **large** for the  $p$  involved, the binomial coefficient would exceed 1,000,000 and then PROGRESS switches to a random selection of  $p$ -subsets. It is possible for the user to preset the number of  $p$ -subsets to be considered. The more  $p$ -subsets you take, the lower the objective function will be, but at the cost of more computation time. On the other hand, one must select enough  $p$ -subsets for the probability of drawing at least one uncontaminated  $p$ -subset to be close to 1 (otherwise, the fit could be based on bad

observations only). In (Rousseeuw and Leroy, 1987, page 198) this minimal number of  $p$ -subsets is expressed in function of the number of variables and the allowed percentage of contamination. The default number of subsets drawn in PROGRESS can be found in Table 4. For  $p \leq 9$  these numbers exceed the required minimum, whereas for larger  $p$  the default is fixed at 3000 subsets so as to avoid extremely long calculations. But as already mentioned, the user can always modify the proposed number of  $p$ -subsets.

Finally, for all **intermediate** values of  $n$  the user can choose between considering all  $p$ -subsets or drawing a certain number of random  $p$ -subsets. As always, the program applies default choices unless the user explicitly asks to override them.

|                                                      | <b>p</b> |      |      |      |      |      |      |      |      |      |
|------------------------------------------------------|----------|------|------|------|------|------|------|------|------|------|
|                                                      | 1        | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   |
| <b>n is 'small' if <math>n \leq</math></b>           | 500      | 50   | 22   | 17   | 15   | 14   | 0    | 0    | 0    | 0    |
| <b>n is 'large' if <math>n \geq</math></b>           | $10^6$   | 1414 | 182  | 71   | 43   | 32   | 27   | 24   | 23   | 22   |
| <b>default number of <math>p</math>-subsets used</b> | 500      | 1000 | 1500 | 2000 | 2500 | 3000 | 3000 | 3000 | 3000 | 3000 |

Table 4: Sample sizes  $n$  which are considered to be small or large (for a given  $p$ ). Also the default number of  $p$ -subsets used in PROGRESS is listed.

## 2. compute hyperplane through these $p$ observations

If the  $\mathbf{x}_i$  are in *general position* then every  $p$ -subset determines a unique hyperplane, that is found by solving the linear system formed by these  $p$  observations.

In practice also a singular  $p$ -subset can occur, and then PROGRESS draws a new  $p$ -subset. The output then reports the total number of singular  $p$ -subsets that were encountered.

## 3. if regression with intercept $\Rightarrow$ adjust intercept

Here, ‘intercept adjustment’ stands for a technique which decreases the objective value of a given fit. We will apply it to each  $p$ -subset. After the hyperplane through the  $p$  observations is determined, we have an initial estimate of the slope and the intercept, given by  $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_{p-1}, \tilde{\theta}_p)$  where  $\tilde{\theta}_p$  is the intercept. The corresponding objective value for LQS then equals

$$|r(\tilde{\theta})|_{h:n} = |y_j - x_{j1}\tilde{\theta}_1 - \dots - x_{j,p-1}\tilde{\theta}_{p-1} - \tilde{\theta}_p|_{h:n}. \quad (7)$$

For LTS we can rewrite (3) accordingly. The adjusted intercept  $\tilde{\theta}'_p$  is then defined as the LQS (resp. LTS) location estimate applied to the univariate

data set  $\{t_i = y_i - x_{i1}\tilde{\theta}_1 - \dots - x_{i,p-1}\tilde{\theta}_{p-1}; i = 1, \dots, n\}$ , i.e.

$$\tilde{\theta}'_p = \underset{\mu}{\operatorname{argmin}} |t_j - \mu|_{h:n} \quad (8)$$

for LQS. By construction, (8) yields a lower objective value than (7). In simple regression ( $p = 2$ ), it follows from (Steele and Steiger, 1986) that if all 2-subsets are used and their intercept is adjusted each time, we obtain the exact LQS.

As indicated in Remark 1, the LQS and LTS location estimates can be found by an explicit algorithm. For LQS it is the midpoint of the shortest interval that contains  $h$  observations, as was proved in (Rousseeuw, 1984, page 873). We thus have to order the univariate observations  $\{t_1, \dots, t_n\}$  to  $t_{1:n} \leq \dots \leq t_{n:n}$  and then compute the length of the contiguous intervals that contain  $h$  points. When the smallest length is attained by several intervals, we take the median of the corresponding midpoints.

The univariate LTS estimator corresponds to the mean of the subset that contains  $h$  observations and that has the smallest sum of squares. This sum of squares is defined as the sum of the squared deviations from the subset mean: given an  $h$ -subset  $t_{i:n}, \dots, t_{i+h-1:n}$  with mean  $\bar{t}^{(i)}$  we have

$$SQ^{(i)} = \sum_{j=i}^{i+h-1} (t_{j:n} - \bar{t}^{(i)})^2.$$

Note that the selected  $h$ -subset has to consist of successive observations, which is why we had to order the  $t_1, \dots, t_n$  first.

For a recent study of the effect of intercept adjustment on the performance of LTS regression, see Croux et al. (1996).

In order to adjust the intercepts the univariate LQS and LTS methods were included into PROGRESS, which also allows the user to analyze data sets that were univariate from the start. As in the regression situation, the preliminary scale is then defined by (4) resp. (5), both of which come out of the univariate algorithms. For LQS it is half the length of the shortest interval, whereas for LTS it is the square root of the smallest sum of squares divided by  $h$ . We then obtain the final scale estimate as in (6).

## 5 Coefficient of determination ( $R^2$ )

Let us first consider the regression model *with* intercept. Along with the classical least squares (LS) comes the coefficient of determination, which measures the proportion of the variance of the response variable explained by the linear model, i.e.

$$0 \leq R^2 = \frac{\operatorname{Var}(y_i) - \operatorname{Var}(r_i)}{\operatorname{Var}(y_i)} = 1 - \frac{\operatorname{Var}(r_i)}{\operatorname{Var}(y_i)} \leq 1. \quad (9)$$

The denominator in this expression measures the variability of the response in a model without explanatory variables, which in this case is the univariate model  $y_i = \mu + \sigma e_i$ . If we denote the LS coefficient estimate of a sample  $(X, \mathbf{y})$  by  $\hat{\theta}_{LS}(X, \mathbf{y})$ , and use the scale estimate given by

$$s_{LS}^2(X, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \mathbf{x}_i \hat{\theta}_{LS}(X, \mathbf{y}))^2, \quad (10)$$

we can rewrite (9) as

$$R_{LS}^2 = 1 - \frac{s_{LS}^2(X, \mathbf{y})}{s_{LS}^2(\mathbf{1}, \mathbf{y})}. \quad (11)$$

By analogy, we propose the robust counterpart given by

$$R_{LQ(T)S}^2 := 1 - \frac{s_{LQ(T)S}^2(X, \mathbf{y})}{s_{LQ(T)S}^2(\mathbf{1}, \mathbf{y})}. \quad (12)$$

Note that when using definition (12), the robust coefficient of determination always falls in the interval  $[0, 1]$ . This was not guaranteed by the earlier version of  $R^2$  defined in (Rousseeuw and Leroy, 1987, page 44) and implemented in the first version of PROGRESS. There the denominator  $(\text{mady})^2 = (\text{med}_i |y_i - \text{med}_j y_j|)^2$  was used, whereas now  $s_{LQS}(\mathbf{1}, \mathbf{y}) = |y - \hat{\theta}_{LQS}(\mathbf{1}, \mathbf{y})|_{h:n}$  which is just the scale estimate of the univariate LQS applied to the response.

An analogous reasoning works for the regression model without intercept. In that case, the model without explanatory variables reduces to  $y_i = \sigma e_i$  without any location parameter. For LS,

$$R_{LS}^2 = 1 - \frac{s_{LS}^2(X, \mathbf{y})}{s_{LS}^2(\mathbf{0}, \mathbf{y})} = 1 - \frac{\sum_i (y_i - \mathbf{x}_i \hat{\theta}_{LS})^2}{\sum_i y_i^2} \quad (13)$$

and we propose the following robust counterparts:

$$R_{LQS}^2 = 1 - \frac{s_{LQS}^2(X, \mathbf{y})}{s_{LQS}^2(\mathbf{0}, \mathbf{y})} = 1 - \frac{((y - \mathbf{x} \hat{\theta}_{LQS})^2)_{h:n}}{(y^2)_{h:n}}$$

and

$$R_{LTS}^2 = 1 - \frac{s_{LTS}^2(X, \mathbf{y})}{s_{LTS}^2(\mathbf{0}, \mathbf{y})} = 1 - \frac{\sum_{i=1}^h ((y - \mathbf{x} \hat{\theta}_{LTS})^2)_{i:n}}{\sum_{i=1}^h (y^2)_{i:n}}.$$

## 6 Diagnostics

Observations in regression data essentially belong to four types:

- regular observations* with internal  $\mathbf{x}_i$  and well-fitting  $y_i$ ,
- vertical outliers* with internal  $\mathbf{x}_i$  and non-fitting  $y_i$ ,
- good leverage points* with outlying  $\mathbf{x}_i$  and well-fitting  $y_i$ ,
- bad leverage points* with outlying  $\mathbf{x}_i$  and non-fitting  $y_i$ .

Figure 1 shows these four types in simple regression. Regression diagnostics aim to detect observations of one or more of these types. Here we will consider three robust diagnostics: standardized residuals, the resistant diagnostic, and the diagnostic plot.

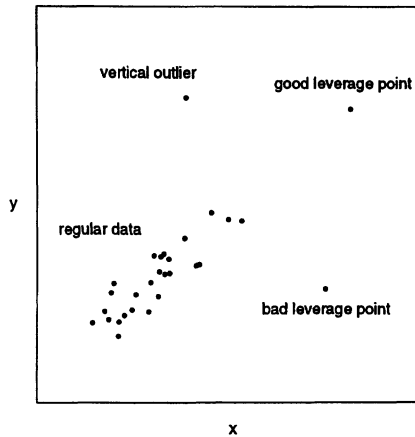


Figure 1: Simple regression data with points of all four types.

1. **Standardized residuals** are defined as  $r_i(\theta)/s(\theta)$  where  $s(\theta)$  denotes a robust scale estimate based on the residuals. Here we will use  $s_{LQS}(\theta) = c_{h,n}|r(\theta)|_{h:n}$  or  $s_{LTS}(\theta) = d_{h,n}\sqrt{\frac{1}{h}\sum_1^h r^2(\theta)_{i:n}}$ . Standardized residuals help us to distinguish between well-fitting and non-fitting observations by comparing their absolute values to some yardstick, e.g.

compare  $|r_i(\theta)|/s(\theta)$  to 2.5.

We use the yardstick 2.5 since it would determine a (roughly) 99% tolerance interval for the  $e_i$  if they had a standard gaussian distribution. Since the standardized residuals approximate the  $e_i$ , we will consider an observation as non-fitting if its standardized residual lies (far) outside this tolerance region.

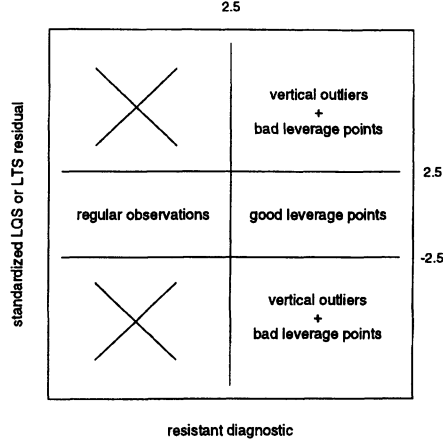


Figure 2: Classification of observations by plotting the standardized residuals versus their resistant diagnostics.

**2. The resistant diagnostic.** Non-regular observations have the property that they are ‘far away’ from some hyperplane in  $\mathbb{R}^{p+1}$  (that is, further away than the *majority* of the observations). The vertical outliers and the bad leverage points are clearly far away from the ideal regression plane given by  $y = \mathbf{x}^t \theta$ . But also a good leverage point lies far away, relative to some other hyperplane that goes through the center of the regular observations. To define the ‘distance’ of an observation  $(\mathbf{x}_i, y_i)$  to a plane  $y = \mathbf{x}^t \theta$  we can use its absolute standardized residual. If we now define

$$U_i = \sup_{\theta} \frac{|r_i(\theta)|}{s_{LQS}(\theta)} \quad \text{or} \quad U_i = \sup_{\theta} \frac{|r_i(\theta)|}{s_{LTS}(\theta)}$$

we expect outliers to have a large  $U_i$ . Since the  $U_i$  are difficult to compute exactly, we approximate them by taking the maximum over all  $\tilde{\theta}$  that are computed inside the LQS/LTS algorithm. For each observation, this yields the value

$$u_i = \max_{\tilde{\theta}} \frac{|r_i(\tilde{\theta})|}{s_{LQS}(\tilde{\theta})} \quad \text{or} \quad u_i = \max_{\tilde{\theta}} \frac{|r_i(\tilde{\theta})|}{s_{LTS}(\tilde{\theta})}. \quad (14)$$

Therefore we only need to store one array  $(u_1, \dots, u_n)$  that has to be updated at each  $p$ -subset. Finally, we define the resistant diagnostic for each observation by standardizing the  $u_i$ , yielding

$$\text{resistant diagnostic}_i = \frac{u_i}{\text{med}_{j=1, \dots, n} u_j}. \quad (15)$$



In the new version of PROGRESS the resistant diagnostic is available for both LQS and LTS, and it is based on the trial estimates  $\tilde{\theta}$  after location adjustment. From (14) it is clear that non-regular observations will have a large  $v_i$  and consequently a large resistant diagnostic. Simulations have indicated that we may consider (15) as ‘large’ if it exceeds 2.5. Combining the standardized residuals with the resistant diagnostic leads to the diagram in Figure 2. However, a disadvantage of Figure 2 is that it cannot distinguish between vertical outliers and bad leverage points.

**3. The diagnostic plot** makes the complete classification into the four types. Since leverage points are outlying in the space of the regressors  $\mathbf{x}_i$ , one can distinguish them from vertical outliers by analyzing their  $\mathbf{x}$ -components. For this we can run MINVOL on  $X = \{\mathbf{x}_i; 1 \leq i \leq n\}$ . This program computes the Minimum Volume Ellipsoid (MVE) location estimate  $T(X)$  and scatter matrix  $C(X)$ . The MVE is a highly robust estimator of location and scatter, introduced by Rousseeuw (1985). The corresponding *robust distance* of an observation to the center is then given by

$$RD(\mathbf{x}_i) := \sqrt{(\mathbf{x}_i - T(X))^t C(X)^{-1} (\mathbf{x}_i - T(X))}.$$

Since the squares of these distances roughly have a chi-squared distribution when there are no outliers among the  $\mathbf{x}_i$ , we will classify an observation as a leverage point if its  $RD(\mathbf{x}_i)$  exceeds the cutoff value  $\sqrt{\chi_{p,0.975}^2}$ . If we combine this information with the standardized LQS or LTS residual, we obtain the diagnostic plot of (Rousseeuw and van Zomeren, 1990) shown in Figure 3.

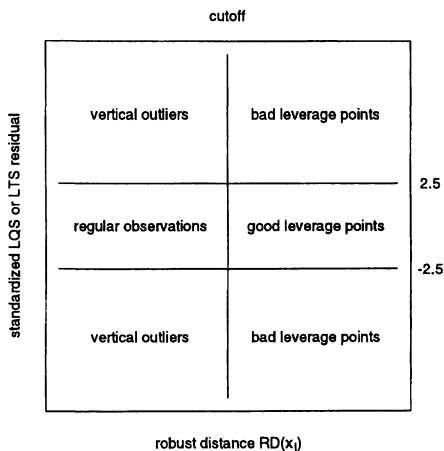


Figure 3: Diagnostic plot, obtained by plotting the standardized robust residuals versus the robust distances  $RD(\mathbf{x}_i)$ .

## 7 Software availability

The programs PROGRESS and MINVOL can be obtained from our website <http://win-www.uia.ac.be/u/statis>

Questions or remarks about the implementation can be directed to [Mia.Hubert@uia.ua.ac.be](mailto:Mia.Hubert@uia.ua.ac.be).

The LMS, LTS and MVE methods are also available in S-PLUS as the functions `lmsreg`, `ltsreg` and `cov.mve`. Moreover, the functions LMS, LTS and MVE based on the recent versions of PROGRESS and MINVOL have been incorporated into SAS/IML (Version 6.12) in 1996. Their documentation can be obtained by writing to [sasaxs@unx.sas.com](mailto:sasaxs@unx.sas.com).

## Appendix: Proof of Theorem 1

We show that the proof of the breakdown value of the LMS (Rousseeuw, 1984, page 878) remains valid after making the necessary modifications. The proofs for LQS and LTS are very similar, so we will mainly consider the LQS estimator.

First suppose  $h < \lfloor \frac{n+p+1}{2} \rfloor$ . We obtain the lower bound on  $\varepsilon_n^*$  by replacing  $h - p + 1 - 1 = h - p$  observations of  $Z$ , yielding  $Z'$ . Define  $\theta = \hat{\theta}_{LQS}(Z)$  and  $\theta' = \hat{\theta}_{LQS}(Z')$ . The  $n - h + p \geq h$  original points  $(\mathbf{x}_i, y_i)$  then satisfy  $|r_i(\theta)| = |y_i - \mathbf{x}_i\theta| \leq M = \max_Z |r_i(\theta)|$  such that for the corrupted data set  $Z'$ ,  $|r_i(\theta')|_{h:n} \leq |r_i(\theta)|_{h:n} \leq M$ . For  $p > 1$  we refer to the geometrical construction of (Rousseeuw, 1984). In his notation, the set  $Z' \setminus A$  contains at most  $n - (n - h + p - (p - 1)) = h - 1$  observations. If we assume  $\|\theta' - \theta\| \geq 2(\|\theta\| + M/\rho)$ , this implies

$$|r_i(\theta')|_{h:n} > M, \quad (16)$$

a contradiction. Therefore,  $\|\theta'\|$  remains bounded. For  $p = 1$  we set  $C = 2M/N = 2M/\min_Z |x_i|$ . Now suppose  $|\theta - \theta'| > C$ . For all non-contaminated observations we have that  $|r_i(\theta) - r_i(\theta')| = |y_i - x_i\theta - y_i - x_i\theta'| = |x_i||\theta - \theta'| > NC = 2M$ , from which we get  $|r_i(\theta')| \geq |r_i(\theta) - r_i(\theta')| - |r_i(\theta)| > 2M - M = M$ . Again this implies (16) and thus a bounded  $\theta'$ . The upper bound on  $\varepsilon_n^*$  follows from the fact that we can put  $h - p + 1$  bad observations on a hyperplane that contains  $p - 1$  original points. Then  $(h - p + 1) + (p - 1) = h$  observations satisfy  $y'_i = \mathbf{x}'_i\theta'$ , and thus  $\theta' = \hat{\theta}_{LQS}(Z')$ . Making the hyperplane steeper will break down the estimator.

For  $h \geq \lfloor \frac{n+p+1}{2} \rfloor$ , we obtain the lower bound analogously to the previous case. Just observe that we now have  $n - (n - h + 1) + 1 = h$  original observations, and that  $Z' \setminus A$  has at most  $n - (h - (p - 1)) = n - h + p - 1 \leq h - 1$  points. The remaining inequality  $\varepsilon_n^* \leq (n - h + 1)/n$  can

be proved as follows. Take some  $M > \|\theta\|$ . Then we show that we can always construct a corrupted sample  $Z'$  with  $n - h + 1$  bad observations, such that  $\|\theta'\| = \|\hat{\theta}_{LQS}(Z')\| \geq M$ . Letting  $M$  go to infinity will then cause the LQS to break down. Define  $M_X = \max_i \|\mathbf{x}_i\|$ . Now we set all the  $n - h + 1$  replaced observations equal to the point  $(\mathbf{x}, y) = (\mathbf{x}, 2M_X M + K)$  for which  $\|\mathbf{x}\| = M_X$  and  $K > 0$ . These replaced observations satisfy  $|\mathbf{x}_i \theta| \leq \|\mathbf{x}_i\| \|\theta\| < \|\mathbf{x}\| M = M_X M < y$  and thus  $|r_i(\theta)| = |y_i - \mathbf{x}_i \theta| \geq |y| - |\mathbf{x} \theta| > M_X M + K$ . As  $n - h + 1 > n - h$  this yields  $|r_i(\theta)|_{h:n} > M_X M + K$ . Since we can choose  $K$  arbitrarily large, the minimum of the objective function of LQS will not be reached for  $\|\theta\| < M$ . Consequently  $\|\theta'\| = \|\hat{\theta}_{LQS}(Z')\|$  has to be larger than  $M$ , which ends the proof. Finally we note that using the same construction, the objective function of LTS satisfies

$$\sum_{i=1}^h (r^2(\theta))_{i:n} > (M_X M + K)^2$$

yielding the same result.

## References

- [1] Croux, C., Haesbroeck, G., and Rousseeuw, P.J. (1996). Location adjustment for the minimum volume ellipsoid estimator. Submitted for publication.
- [2] Donoho, D.L., and Huber, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich Lehmann*, Eds. P. Bickel, K. Doksum, and J.L. Hodges. California: Wadsworth.
- [3] Rousseeuw, P.J. (1984). Least median of squares regression. *J. Am. Statist. Assoc.* **79** 871-880.
- [4] Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, Eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Vol. B, pp. 283-297. Dordrecht: Reidel.
- [5] Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
- [6] Rousseeuw, P.J., and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *J. Am. Statist. Assoc.* **85** 633-639.
- [7] Steele, J.M., and Steiger, W.L. (1986). Algorithms and complexity for least median of squares regression. *Discrete Applied Mathematics* **14** 93-100.

# Dimension reduction via parametric inverse regression

Efstathia Bura

*The George Washington University, Washington, USA*

*Abstract:* In this paper, a linear subspace containing part or all of the information for the regression of a  $m$ -vector  $Y$  on a  $p$ -vector  $X$  and its dimension are estimated via the means of inverse regression. Smooth parametric curves are fitted to the  $p$  inverse regressions through a multivariate linear model, without imposing any strict assumptions on the error distribution. This method is expected to be more powerful in reducing the dimension of a regression problem when compared to *SIR*, the estimation procedure proposed by Li (1991), that is based on fitting piecewise constant functions to the inverse regression curves.

*Key words:* Dimension reduction, regression, linear subspace estimation.

AMS subject classification: 62A99, 62H05.

## 1 Introduction

Let  $Y \in \mathbb{R}^m$  and  $X \in \mathbb{R}^p$  with joint cumulative distribution function (c.d.f.)  $F(Y, X)$ . In a regression setting the behavior of the conditional cumulative distribution function of  $Y$  given  $X$ ,  $F(Y|X)$ , as the value of  $X$  varies in its marginal sample space is under study. As a means of characterizing the regression structure, consider replacing  $X$  by  $k \leq p$  linear combinations of its components,  $\eta_1^T X, \dots, \eta_k^T X$ , without losing information on  $F(Y|X)$  so that, for all values of  $X$ ,

$$F(Y|X) = F(Y|\eta_1^T X, \dots, \eta_k^T X) = F(Y|\eta^T X) \quad (1)$$

where  $\eta$  is the  $p \times k$  matrix with columns  $\eta_j$ , and  $F(\cdot|\cdot)$  denotes the conditional c.d.f. of the first argument given the second. Equation (1) holds trivially when  $\eta = I_p$ , where  $I_p$  denotes the identity matrix of dimension

$p$ , and thus it imposes no restrictions on  $F(Y|X)$ . It can be expressed equivalently as

$$Y \perp\!\!\!\perp X | \eta^T X \quad (2)$$

where the notation  $U \perp\!\!\!\perp V | W$  in (2) means that  $U$  is independent of  $V$  given any value for  $W$  (Dawid, 1979). Both (1) and (2) express the fact that the conditional c.d.f. of  $Y|X$  depends on  $X$  only through  $\eta^T X$ , the coordinates of a projection of  $X$  onto the  $k$ -dimensional linear subspace spanned by the columns of  $\eta$ . Consequently,  $\eta^T X$  can be used in place of  $X$  without loss of information on the regression.

An example where (2) holds is the additive-error regression model

$$Y|X = g(\eta_1^T X, \dots, \eta_k^T X) + \epsilon$$

where  $\epsilon \perp\!\!\!\perp X$  and  $E(\epsilon) = 0$ .

For any vector or matrix  $\alpha$ , let  $S(\alpha)$  denote its range space and  $\dim(S(\alpha))$  denote its dimension. If (1) holds then it also holds with  $\eta$  replaced by any basis for  $S(\eta)$ . In this sense, (1) and (2) can be regarded as statements about  $S(\eta)$  rather than statements about  $\eta$  per se. Thus, when (2) holds we follow Li (1991, 1992) and call  $S(\eta)$  a *dimension-reduction subspace* for  $F(Y|X)$  or for the regression of  $Y$  on  $X$ .

Obviously, the smallest dimension-reduction subspace provides the greatest dimension reduction in the predictor vector. Unfortunately, smallest or *minimum* dimension-reduction subspaces (Cook 1994a) are not always unique. To circumvent the latter, Cook (1994b, 1996) introduced the notion of *central* dimension-reduction subspaces:

**Definition 1** A subspace  $S$  is a central dimension-reduction subspace for the regression of  $Y$  on  $X$  if (a)  $S$  is a dimension-reduction subspace and (b)  $S \subset S_{drs}$  for all dimension-reduction subspaces  $S_{drs}$ , i.e.  $S = \cap S_{drs}$ . A central dimension-reduction subspace will be denoted by  $S_{Y|X}(\cdot)$ .

The intersection of all dimension-reduction subspaces  $\cap S_{drs}$  is trivially a subspace but it is not necessarily a dimension-reduction one. Also, it is easy to see that a central dimension-reduction subspace is a minimum dimension-reduction subspace but the converse is not always true. In fact, there are regression problems for which the central dimension-reduction subspace does not exist. A detailed discussion of these issues can be found in Cook (1994b, 1996).

By definition, a central dimension-reduction subspace, being the intersection of all dimension-reduction subspaces, is unique when it exists. The existence of central subspaces can be assured by placing fairly weak restrictions on aspects of the joint distribution of  $Y$  and  $X$  (Cook 1994a, 1996). In

this paper, we concentrate on regressions where central dimension-reduction spaces exist.

The subspace  $S_{Y|X}$  is considered a “super-parameter” that is used to index the conditional distribution of  $Y$  given  $X$  and its estimation is the main theme of this work. Throughout the rest of this article, the columns of the  $p \times k$  matrix  $\eta$  form a basis for the central space  $S_{Y|X}$ , and  $k$  is used to denote its dimension.

### 1.1 Inverse regression and $SIR$

Methods are available for estimating portions of the central subspace  $S_{Y|X}$  if we are willing to place certain conditions on the marginal distribution of the predictors. The method that will be presented in this article is based on inverse regression.

Let  $S_{E(X|Y)}$  denote the subspace spanned by  $\{E(X|Y) - E(X) : Y \in \Omega_Y\}$ , where  $\Omega_Y \subset \mathbb{R}^m$  is the marginal sample space of  $Y$ . The condition that the marginal distribution of the predictors  $X$  must satisfy in order for inverse regression to be useful in estimating a portion of the central subspace is stated in the following theorem. The theorem, as presented by Li (1991), is based on an arbitrary dimension-reduction subspace which need not be central. However, the version here is stated in terms of the central subspace. Throughout this article, boldface capital Latin letters will denote matrices, even though other symbols will also be used for the same purpose provided there is no fear of confusion.

**Theorem 1** *Assume that the central subspace  $S_{Y|X}(\eta)$  exists for  $F(Y|X)$ , and that, for all  $b \in \mathbb{R}^p$ ,  $E(b^T X | \eta^T X)$  is linear in  $\eta^T X$ . Then the centered inverse regression curve  $E(X|Y) - E(X)$  satisfies*

$$E(X|Y) - E(X) \in S(\Sigma_x \eta)$$

*Equivalently,*

$$S_{E(X|Y)} \subset S(\Sigma_x \eta) = \Sigma_x S_{Y|X}$$

*where  $\Sigma_x = \text{Cov}(X)$ .*

**Proof:** Li (1991) proved Theorem 1 for any dimension reduction subspace  $S(\eta)$  so that (2) is satisfied. It is obvious that if the theorem holds for an arbitrary dimension reduction subspace, it also holds for the intersection of all dimension reduction subspaces; that is, the central dimension reduction subspace  $S_{Y|X}$ , provided it exists.  $\square$

The linearity condition on  $E(b^T X | \eta^T X)$  is required to hold only for the basis  $\eta$  of the central subspace.  $\eta$  being unknown, in practice we may

require that it hold for all possible  $\eta$ , which is equivalent to elliptical symmetry of the distribution of  $X$  (Eaton, 1986). Li (1991) mentioned that the linearity condition is not a severe restriction, since most low-dimensional projections of a high-dimensional data cloud are close to being normal (Diaconis and Freedman, 1984; and Hall and Li, 1993). In addition, there often exist transformations of the predictors that make them comply with the linearity condition. Cook and Nachtsheim (1994) suggested re-weighting of the predictor vector to make it elliptically contoured.

In the next corollary, which follows directly from Theorem 1, the analogous result is given for a standard random vector. Suppose that  $\Sigma_x > 0$  and let  $Z$  be the standardized version of  $X$ ,

$$Z = \Sigma_x^{-1/2}(X - E(X))$$

Obviously,  $E(Z) = 0$  and  $\text{Cov}(Z) = I_p$ . Also, since  $Z$  is a 1-1 and onto linear transformation of  $X$ ,  $Y \perp\!\!\!\perp X | \eta^T X$  if and only if  $Y \perp\!\!\!\perp Z | \beta^T Z$ , where  $\beta = \Sigma_x^{1/2}\eta$  or  $\beta_i = \Sigma_x^{1/2}\eta_i$ ,  $i = 1, 2, \dots, k$ .

### Corollary 1

$$E(Z|Y) \in S(\Sigma_x^{1/2}\eta) = S(\beta) = S_{Y|Z}$$

This corollary readily implies that  $E(Z|Y) = \mathbf{P}_\beta E(Z|Y)$ , where  $\mathbf{P}_\beta$  is the orthogonal projection operator for  $S(\beta)$  with respect to the usual inner product.

Corollary 1 also implies that  $S_{E(Z|Y)}$  is a subspace of  $S_{Y|Z}$ . This does not guarantee equality between  $S_{E(Z|Y)}$  and  $S_{Y|Z}$ , and thus, inference about  $S_{E(Z|Y)}$  possibly covers only part of  $S_{Y|Z}$ . For example, if  $Y = Z_1^2$ , with  $Z_1$  being the first coordinate variable of  $Z$ , and if  $Z_1$  is symmetric about its mean, then  $E(Z|Y) = 0$  even though  $S_{Y|Z} = \text{span}((1, 0)^T)$ . For a broader discussion of the inability of *SIR*, and consequently of the method developed in this paper, to diagnose this symmetric dependence see Cook and Weisberg (1991). The missed part of  $S_{Y|Z}$  might be recovered from higher order moments of the conditional distribution of  $Z$  given  $Y$  (Cook and Weisberg 1991; Li 1992), but such issues are not addressed in this article. We assume throughout that  $S_{E(Z|Y)}$  is non-trivial, in the sense that it contains non-zero directions, should they exist.

Theorem 1 and Corollary 1 lead to the use of inverse regression as an estimation means of part or possibly the whole of the central dimension-reduction subspace. One such method is *SIR* (Sliced Inverse Regression), proposed by Li (1991). In *SIR*, the range of the one-dimensional variable  $Y$  is partitioned into a fixed number of slices and the  $p$  components

of  $Z$  are regressed on  $\tilde{Y}$ , a discrete version of  $Y$  resulting from slicing its range, giving  $p$  one-dimensional regression problems, instead of the possibly high-dimensional forward regression of  $Y$  on  $Z$ . Then, a rather crude nonparametric estimate of the inverse curve  $E(X|Y)$  serves to estimate the central dimension-reduction subspace. *SIR* includes an asymptotic test for inferring about  $d$ , a lower bound on  $k$ .

But, even though Li (1991) introduced an innovative way of reducing the dimension in a regression problem, *SIR* has limitations of which the most important is that *SIR* can be ambiguous about the estimate of the dimension as the latter depends sometimes crucially on the choice of the number of slices. This can be easily avoided by using standard regression estimation techniques.

In this article, smooth parametric curves are fitted to the  $p$  inverse regressions in order to estimate the central subspace  $S_{Y|X}(\eta)$ , without imposing any restrictions on the dimension of the response vector  $Y$ .

## 2 Parametric inverse regression

For simplicity assume that  $X$  is standardized to have 0 mean and the identity covariance matrix. To model the conditional expectation of  $X$  given  $Y$ , a multivariate linear model is fitted with  $X$  being the response,  $X^T = (x_1, \dots, x_p)$ , and  $Y$ ,  $Y^T = (y_1, \dots, y_m)$ , the explanatory vector. Let

$$E \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}^T = \begin{bmatrix} f_1(Y) & \cdots & f_q(Y) \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix}$$

where the  $f_i$ 's are arbitrary,  $\mathbb{R}$ -valued linearly independent known functions of  $Y$ . Suppose that a random sample of size  $n$  is available on  $(Y, X)$ . Then, including a matrix of errors  $\mathbf{E}_n$ , the model becomes

$$\mathbf{X}_n|Y = \mathbf{Z}_n\mathbf{B} + \mathbf{E}_n \quad (3)$$

where  $\mathbf{X}_n = (x_{ij})$ , a  $n \times p$  random matrix,  $\mathbf{Z}_n = (f_{il})$ , a  $n \times q$  fixed matrix with  $f_{il} = f_l(Y_i)$ , and  $\mathbf{B} = (\beta_{lj})$ , the  $q \times p$  matrix of coefficients. The error matrix  $\mathbf{E}_n$  satisfies

$$E(\mathbf{E}_n|Y) = 0 \quad \text{and} \quad \text{Cov}(\vec{\mathbf{E}}_n|Y) = \Sigma_{x|y} \otimes I_n$$

where  $\Sigma_{x|y}$  is a  $p \times p$  positive definite, unknown matrix, that does not depend on  $Y$ .  $\mathbf{X}_n$ ,  $\mathbf{Z}_n$ , and  $\mathbf{E}_n$  are indexed by the sample size  $n$  to indicate



their dependence upon it. The symbol  $\otimes$  denotes the Kronecker product. Clearly, the rank of  $\mathbf{Z}_n$  is  $q$ . We assume that  $n \geq p$  in order to avoid trivial cases. No distributional assumptions on the errors are made except that the rows  $\epsilon_i^{(n)T}$  of the error matrix  $\mathbf{E}_n$  are independent with mean 0 and constant covariance matrix  $\Sigma_{x|y}$ .

According to (3),  $S_{E(X|Y)}$  is the linear subspace of  $S_{Y|X}$  which is spanned by the rows of  $\mathbf{Z}_n \mathbf{B}$ ; that is  $S(\mathbf{B}^T \mathbf{Z}_n^T) = S_{E(X|Y)}$ . Therefore, since  $\text{rank}(\mathbf{B}^T \mathbf{Z}_n^T) = \text{rank}(\mathbf{Z}_n \mathbf{B})$ ,  $\text{rank}(\mathbf{Z}_n \mathbf{B}) \leq \dim(S_{Y|X})$ , and the  $\text{rank}(\mathbf{Z}_n \mathbf{B})$  is a lower bound on the dimension of the central dimension-reduction subspace.

But,

$$\text{rank}(\mathbf{Z}_n \mathbf{B}) = \text{rank}(\mathbf{B}^T \mathbf{Z}_n^T \mathbf{Z}_n \mathbf{B}) = \text{rank}(\mathbf{B})$$

since  $\mathbf{Z}_n^T \mathbf{Z}_n$  is a positive definite matrix (see [A4.4], Seber, 1977). Thus, the rank of  $\mathbf{Z}_n \mathbf{B}$  is actually equal to the rank of  $\mathbf{B}$ , and hence inference on the dimension of  $S_{E(X|Y)}$  can be based solely on  $\mathbf{B}$  in the sense that an estimate of the rank of  $\mathbf{B}$  constitutes an estimate of a lower bound on the dimension of  $S_{Y|X}$ .

The estimate of  $\mathbf{B}$  to be used for inference on the rank of  $\mathbf{B}$  is the ordinary least squares estimate, given by

$$\hat{\mathbf{B}}_n = (\mathbf{Z}_n^T \mathbf{Z}_n)^{-1} \mathbf{Z}_n^T \mathbf{X}_n \quad (4)$$

### 3 The asymptotic distribution of $\mathbf{B}_n$

Let  $e_i^{(n)}$  be the  $n$ -vector with 1 in the  $i$ th place and zeroes elsewhere. We are interested in the asymptotic distribution of  $\sqrt{n} (\hat{\mathbf{B}}_n - \mathbf{B})$ .

Let  $\mathbf{H}_n$  denote the covariance matrix of  $\sqrt{n} (\hat{\mathbf{B}}_n - \mathbf{B})$ ,

$$\begin{aligned} \mathbf{H}_n &= \text{Cov}(\sqrt{n} \overrightarrow{(\hat{\mathbf{B}}_n - \mathbf{B})}) \\ &= \Sigma_{x|y} \otimes (\mathbf{Z}_n^T \mathbf{Z}_n / n)^{-1} \end{aligned}$$

The notation  $\|\cdot\|_{\max}$  identifies the norm on the vector space of matrices defined by

$$\|(a_{ij})\|_{\max} = \max_{i,j} |a_{ij}|$$

for a matrix  $= (a_{ij})$ . The following lemma about the asymptotic distribution of  $\sqrt{n} (\hat{\mathbf{B}}_n - \mathbf{B})$  follows readily from Theorem 2.4.3, Bunke and Bunke (1986), and the multivariate version of Slutsky's theorem (see [A 4.19], Bunke and Bunke, 1986).

**Lemma 1** *Let  $\mathcal{M}_{pq}^>$  be the space of all  $pq \times pq$  positive definite matrices and let  $\mathcal{F}$  be the space of distributions of the errors  $\mathbf{E}_n$ . If*

$$\mathbf{H}_n \longrightarrow \mathbf{H} \in \mathcal{M}_{pq}^> \quad (5)$$

then

$$\sqrt{n} \overrightarrow{(\hat{\mathbf{B}}_n - \mathbf{B})} \xrightarrow{\mathcal{D}} N_{qp}(0, \mathbf{H}) \quad (6)$$

provided the following three conditions are satisfied

$$\|(\mathbf{Z}_n^T \mathbf{Z}_n)^{-1} \mathbf{Z}_n^T\|_{\max} = o(n^{-1/2}) \quad (I)$$

$$\sup_{F \in \mathcal{F}} \int_{\|x\| > c} \|x\|^2 dF(x) \longrightarrow 0 \quad \text{as } c \rightarrow \infty \quad (II)$$

$$\inf_{\Sigma \in \mathcal{M}(\mathcal{F})} \lambda_{\min}(\Sigma) \geq r > 0 \quad (III)$$

where

$$\mathcal{M}(\mathcal{F}) = \left\{ \int_{\mathbb{R}^p} x x^T dF(x) : F \in \mathcal{F} \right\} \subset \mathcal{M}_p^>$$

The error distributions that are usually considered satisfy Conditions (II) and (III).

Assume that there exists a matrix  $\mathbf{G} \in \mathcal{M}_q^>$ , so that

$$(\mathbf{Z}_n^T \mathbf{Z}_n / n)^{-1} \longrightarrow \mathbf{G} \quad (7)$$

as  $n \rightarrow \infty$ . Also, assume that a consistent estimate  $\hat{\Sigma}_{x|y}$  is available, as  $\Sigma_{x|y}$  is usually unknown. For instance,  $\hat{\Sigma}_{x|y}$  can be taken to be the matrix of residuals from the regression of  $X$  on  $Y$  divided by either  $n$  or,  $n - \text{rank}(\mathbf{Z}) = n - q$ , the denominator choice that makes  $\hat{\Sigma}_{x|y}$  unbiased for  $\Sigma_{x|y}$  (the proof is omitted). Let

$$\hat{\mathbf{H}}_n = \hat{\Sigma}_{x|y} \otimes (\mathbf{Z}_n^T \mathbf{Z}_n / n)^{-1} \quad (8)$$

Then, if (7) holds,

$$\hat{\mathbf{H}}_n \xrightarrow{p} \mathbf{H} \quad (9)$$

(9) is a direct application of the triangle inequality and the fact that continuous functions of consistent estimates are themselves consistent. These remarks result in the following corollary to Lemma (1).

**Corollary 2** *Suppose Conditions (I), (II), and (III) of Lemma 1 hold. Also assume that  $\hat{\Sigma}_{x|y}$  is a consistent estimate of  $\Sigma_{x|y}$ , and that (7) holds. Then,*

$$\sqrt{n} \hat{\mathbf{H}}_n^{-1/2} \overrightarrow{(\hat{\mathbf{B}}_n - \mathbf{B})} \xrightarrow{\mathcal{D}} N(0, I_{pq}) = N(0, I_p \otimes I_q) \quad (10)$$

**Proof:** Since  $\hat{\Sigma}_{x|y}$  is consistent for  $\Sigma_{x|y}$  and (7) holds, (10) is the result of a direct application of the multivariate version of Slutsky's theorem (see [A 4.19] in Bunke and Bunke, 1986), and of Lemma 1.  $\square$

Let  $d = \dim(S_{E(X|Y)})$ . We have shown that  $d = \text{rank}(\mathbf{B})$  and thus, we can use the least squares estimate of  $\mathbf{B}$ , to estimate the dimension of  $S_{E(X|Y)}$  as follows. Let  $\lambda_j, j = 1, \dots, \min(q, p)$ , be the singular values of  $\mathbf{B}$ . Then,  $d$  is the number of the nonzero singular values of  $\mathbf{B}$ , and inference about  $d$  can be made by testing if

$$\Lambda_d^{(1)} = n \sum_{j=d+1}^{\min(q,p)} \lambda_j^2$$

is equal to zero. We have no direct access to  $\Lambda_d^{(1)}$ , but by observing that the rank of a matrix is not affected when the matrix is multiplied by a nonsingular matrix, the inference on  $d$  can be based on the following test statistic

$$\hat{\Lambda}_d^{(1)} = n \sum_{j=d+1}^{\min(q,p)} \hat{\phi}_j^2 \quad (11)$$

where  $\hat{\phi}_j$  are the singular values of

$$\left(\frac{\mathbf{Z}_n^T \mathbf{Z}_n}{n}\right)^{1/2} \hat{\mathbf{B}}_n \hat{\Sigma}_{x|y}^{-1/2} \quad (12)$$

(12) is used in place of  $\hat{\mathbf{B}}_n$  for convenience, as its asymptotic covariance matrix is the identity. Now, the test is based on the asymptotic distribution of  $\hat{\Lambda}_d^{(1)}$ .

## 4 The asymptotic distribution of $\hat{\Lambda}_d^{(1)}$

Given the asymptotic normality of the least squares estimate of  $\mathbf{B}$ , we can obtain the asymptotic distribution of the singular values of a fixed nonsingular transformation of  $\mathbf{B}$  based on a result about the asymptotic distribution of the singular values of a matrix by Eaton and Tyler (1994).

**Theorem 2** *The asymptotic distribution of  $\hat{\Lambda}_d^{(1)}$  defined in (11) and (12) is  $\chi_{(p-d) \times (q-d)}^2$ .*

**Proof:** Consider the singular value decomposition of  $\mathbf{G}^{-1/2} \mathbf{B} \Sigma_{x|y}^{-1/2}$ , where  $\mathbf{G}$  is the positive definite limit matrix of  $n (\mathbf{Z}_n^T \mathbf{Z}_n)^{-1}$ ,

$$\mathbf{G}^{-1/2} \mathbf{B} \Sigma_{x|y}^{-1/2} = \Gamma_1^T \begin{bmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{bmatrix} \Gamma_2^T$$

$\mathbf{D}$  is a  $d \times d$  diagonal matrix with the positive singular values of  $\mathbf{G}^{-1/2} \mathbf{B} \Sigma_{x|y}^{-1/2}$  along its diagonal. Partition  $\Gamma_1^T = (\Gamma_{11}, \Gamma_{12}) : q \times q$ ,  $\Gamma_{11} : q \times d$ ,  $\Gamma_{12} : q \times (q - d)$ ,

$$\Gamma_2^T = \begin{bmatrix} \Gamma_{21}^T \\ \Gamma_{22}^T \end{bmatrix} : p \times p$$

where  $\Gamma_{21}^T : d \times p$ ,  $\Gamma_{22}^T : (p - d) \times p$ . By the Eaton-Tyler result, the limiting distribution of the smallest  $q - d$  singular values of

$$\sqrt{n} \left( \left( \frac{\mathbf{Z}_n^T \mathbf{Z}_n}{n} \right)^{1/2} \hat{\mathbf{B}}_n \hat{\Sigma}_{x|y}^{-1/2} \right)$$

is the same as the limiting distribution of the singular values of the  $(q - d) \times (p - d)$  matrix

$$\sqrt{n} \mathbf{B}_n = \sqrt{n} (\Gamma_{12}^T \left( \frac{\mathbf{Z}_n^T \mathbf{Z}_n}{n} \right)^{1/2} \hat{\mathbf{B}}_n \hat{\Sigma}_{x|y}^{-1/2} \Gamma_{22})$$

By (10), the asymptotic distribution of  $\sqrt{n} \vec{\mathbf{B}}_n$  is

$$\sqrt{n} (\Gamma_{12}^T \left( \frac{\mathbf{Z}_n^T \mathbf{Z}_n}{n} \right)^{1/2} \hat{\mathbf{B}}_n \hat{\Sigma}_{x|y}^{-1/2} \Gamma_{22}) \xrightarrow{\mathcal{D}} N_{(p-d)(q-d)}(0, I_{p-d} \otimes I_{q-d}) \quad (13)$$

But then,  $\hat{\Lambda}_d^{(1)}$  has the same asymptotic distribution as the sum of the squares of the singular values of  $\sqrt{n} (\Gamma_{12}^T \left( \frac{\mathbf{Z}_n^T \mathbf{Z}_n}{n} \right)^{1/2} \hat{\mathbf{B}}_n \hat{\Sigma}_{x|y}^{-1/2} \Gamma_{22})$  which is  $\chi_{(p-d) \times (q-d)}^2$  by (13).  $\square$

Note that the asymptotic test derived above is equivalent to the usual F-test for testing  $d = 0$  when  $p = 1$ ; that is, when we fit  $q$  functions of  $Y$  on the one-dimensional  $X$  and we test the overall validity of the model by testing the hypothesis  $\beta_{11} = \beta_{21} = \dots = \beta_{q1} = 0$ .

#### 4.1 A summarizing theorem

All the key results discussed and proved in the previous sections are summarized in the following theorem.

**Theorem 3** Assume that  $\mathbf{X}_n|Y = \mathbf{Z}_n \mathbf{B} + \mathbf{E}_n$ , with  $E(\mathbf{E}_n) = 0$ ,  $Cov(\vec{\mathbf{E}}_n) = \Sigma_{x|y} \otimes I_n$ , where  $\mathbf{X}_n : n \times p$ ,  $\mathbf{Z}_n : n \times q$ ,  $\mathbf{B} : q \times p$ , with  $\text{rank}(\mathbf{Z}_n) = q$ . Let  $\hat{\mathbf{B}}_n = (\mathbf{Z}_n^T \mathbf{Z}_n)^{-1} \mathbf{Z}_n^T \mathbf{X}_n$  be the ordinary least squares estimate of  $\mathbf{B}$ . Let  $\hat{\Sigma}_{x|y}$  be a consistent estimate of  $\Sigma_{x|y}$  and  $\mathbf{G}_n^{-1} = \mathbf{Z}_n^T \mathbf{Z}_n / n$ . Assume that  $\mathbf{G}_n \rightarrow \mathbf{G}$  pointwise, where  $\mathbf{G}$  is a  $q \times q$  positive definite matrix. Then,

$$\sqrt{n} (\hat{\Sigma}_{x|y}^{-1/2} \otimes \mathbf{G}_n^{-1/2}) (\hat{\mathbf{B}}_n - \mathbf{B}) \xrightarrow{\mathcal{D}} N_{pq}(0, I_p \otimes I_q)$$

Let  $d = \text{rank}(\mathbf{B})$ . Also, let  $\hat{\phi}_j$ ,  $j = 1, \dots, \min(q, p)$ , be the singular values of  $\mathbf{G}_n^{-1/2} \hat{\mathbf{B}}_n \hat{\Sigma}_{x|y}^{-1/2}$ . Then

$$\hat{\Lambda}_d^{(1)} = n \sum_{j=d+1}^{\min(q,p)} \hat{\phi}_j^2$$

is asymptotically distributed as a  $\chi_{(q-d)(p-d)}^2$  random variable.

**Proof:** Both results are immediate consequences of Corollary 2 and Theorem 2.  $\square$

We can use the asymptotic distribution of  $\hat{\Lambda}_d^{(1)}$  to estimate the rank  $d$  of  $\mathbf{B}$ , or equivalently the dimension of the subspace  $S_{E(X|Y)} \subset S_{Y|X}$ , as follows: Fix  $j$  with  $0 \leq j \leq q$ . Compare  $\hat{\Lambda}_j^{(1)}$  to the quantiles of a  $\chi_{(p-j) \times (q-j)}^2$ ; if it is bigger, conclude that  $d > j$ ; if not, conclude that  $d \leq j$ , and repeat the procedure.

## 5 An example

To illustrate the method, we consider the *Horse Mussel Data*: The data consist of a sample of 201 horse mussel measurements collected in the Marlborough Sounds, which are located off the northeast coast of New Zealand's South Island (Camden, 1989). The response variable is muscle mass  $M$ , the edible portion of the mussel, in grams. The quantitative predictors are shell width  $W$ , shell length  $L$ , in  $mm$ , and shell mass  $S$  in grams. The actual sampling method is unknown, but we assume that the data are i.i.d. observations from the overall mussel population. The  $R$ -code (Cook and Weisberg, 1994) was used for the computations.

In Figure 1a a scatterplot matrix of the response, shell length, shell width and shell mass is presented. It is evident that the linearity condition needed for *SIR* to work may be violated. The transformed variables  $W^{1/2}$  and  $S^{1/4}$  will be used in place of  $W$  and  $S$ , respectively, so that the linearity condition is satisfied by the regressor variables.

Theorem 1 applies to the transformed data and *SIR* can be used to estimate the central dimension-reduction subspace. The results of applying *SIR* to the regression of  $M$  on  $L$ ,  $W^{1/2}$  and  $S^{1/4}$  are given in Tables 1 and 2; Table 1 contains the results when 5 slices were used and Table 2 when 20 slices were used. The rows of both tables summarize hypothesis tests of the form  $d = j$  versus  $d > j$ . For example, the first row gives the statistic  $\hat{\Lambda}_0 = 154.7$  with  $(p-d)(H-d-1) = (3-0)(5-1) = 12$  degrees of freedom and a p-value of 0.000. As it can be seen from the two tables, *SIR* gives contradictory results: it estimates the dimension to be 1 or 2, depending on the number of slices used.

Now, consider fitting smooth parametric curves. The scatterplot matrix in Figure 1b suggests fitting quadratic curves on all three inverse regression plots. The results of the analysis are given in Table 3.

The test indicates a one-dimensional structure supporting that one linear combination of the regressors can be sufficient to characterize the behavior of the conditional c.d.f. of  $M$  given  $L, W^{1/2}, L^{1/4}$ :

$$0.0235275L + 0.003176W^{1/2} + 0.00541707S^{1/4} \tag{14}$$

The same conclusion of one-dimensional structure is also reached using regression graphical techniques to estimate the structural dimension of the this regression problem (see Cook and Weisberg, 1994).

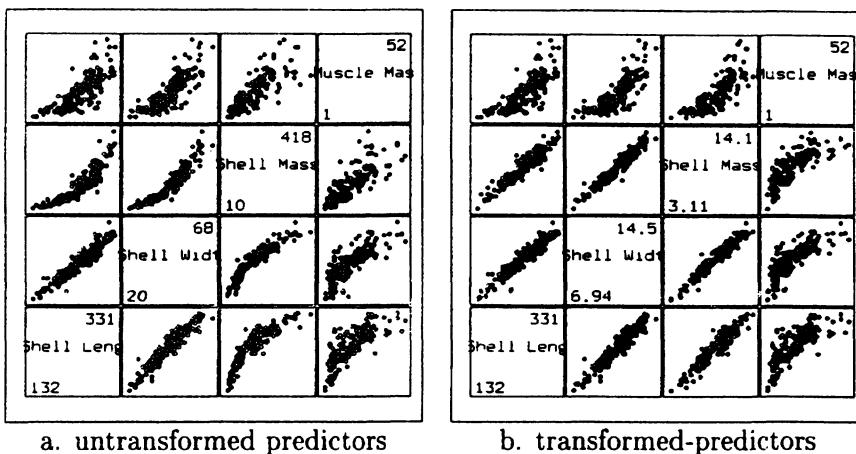


Figure 1: Scatterplots of the Mussel Data

Table 1: SIR results for  $H = 5$

| j | $\hat{\Lambda}_j$ | DF | $p - value$ |
|---|-------------------|----|-------------|
| 0 | 154.7             | 12 | 0.000       |
| 1 | 14.81             | 6  | 0.022       |
| 2 | 4.973             | 2  | 0.083       |

Table 2: SIR results for  $H = 20$

| j | $\hat{\Lambda}_j$ | DF | $p - value$ |
|---|-------------------|----|-------------|
| 0 | 177.2             | 45 | 0.000       |
| 1 | 31.55             | 28 | 0.293       |
| 2 | 9.877             | 13 | 0.704       |

## 6 Discussion

In order to estimate a lower bound on the dimension of the central dimension-reduction subspace  $S_{Y|X}$ , the conditional expectation of the standardized

| j | $\hat{\Lambda}_j^{(1)}$ | DF | $p - value$ |
|---|-------------------------|----|-------------|
| 0 | 502.9                   | 9  | 0.000       |
| 1 | 6.998                   | 4  | 0.136       |
| 2 | 3.8210E-27              | 1  | 1.000       |

Table 3: Parametric results for the Mussel data

$X$  given  $Y$  was modeled according to the linear model (3) placing relaxed conditions on the error distribution, namely zero mean and constant covariance structure. The decision on what model to fit is based on data inspection. The selected model should be a sufficiently complex model that accommodates the data. For example, if polynomials are fitted, the degree should be a number that provides a good fit to all inverse regression curves.

An asymptotic  $\chi^2$  test for the dimension  $d$  of  $S_{E(X|Y)}$  was obtained as a result of the asymptotic normality of the least squares estimate of  $\mathbf{B}$ . The estimated dimension is in fact an estimate of a lower bound for the dimension of  $S_{Y|X}$ .

The  $d$  eigenvectors of the least squares estimate of  $\mathbf{B}$ , that correspond to its  $d$  largest eigenvalues, multiplied by  $\mathbf{Z}_n$  yield estimates of  $d$  of the basis vectors of  $S_{Y|X}$ . They, in turn, can be scaled back to estimates of basis vectors of the central dimension-reduction subspace for the non-standardized  $X$ , by multiplication with  $\hat{\Sigma}_x^{-1/2}$ , where  $\hat{\Sigma}_x$  is the moment estimate of  $\Sigma_x$ .

These results can be extended to the non-constant covariance structure model, under certain conditions. In addition, a similar test has been developed for the case where the inverse regression curves are not all of the same shape. This test does not have an asymptotic distribution with quantiles as easy to compute as these of a  $\chi^2$ . All of the above developments can be found in Bura (1996).

The technique of this article does not suffer from most of the shortcomings of  $SIR$  and requires neither the marginal distribution of  $X$  to be normal nor  $Y$  to be one-dimensional. Further research is needed to assess the sensitivity of the method to outliers. The power of the test is also expected to be higher due to the fitting method.

As an aside, it is worthwhile to comment that even though the estimation procedure developed in this paper was motivated by the use of inverse regression as a means to reduce the dimension of a forward regression problem, it is also a method of estimating the linear subspace spanned by a regression curve. In this context, if the linear subspace is estimated to

be  $\{0\}$  and this cannot be attributed to symmetric dependence (see Cook and Weisberg, 1991), we can possibly infer that the regression curve is intrinsically nonlinear.

## References

- [1] Bunke, H. and Bunke, O. (Editors) (1986). *Statistical Inference in Linear Models. Statistical Methods of Model Building*, Vol. 1. Chichester: Wiley.
- [2] Bura, E. (1996). Dimension reduction via inverse regression. Ph.D. Dissertation, School of Statistics, University of Minnesota.
- [3] Cook, R. D. (1994a). On the interpretation of regression plots. *J. Am. Statist. Assoc.* **89**, 177–189.
- [4] Cook, R. D. (1994b). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proc. of the Section on Physical and Engineering Sciences*, Alexandria, VA: American Statistical Association, pp. 18–25.
- [5] Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Am. Statist. Assoc.* **91**, 983–992.
- [6] Cook, R. D. and Nachtsheim, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *J. Am. Statist. Assoc.* To appear.
- [7] Cook, R. D. and Weisberg, S. (1991). Discussion of “Sliced inverse regression” by K. C. Li. *J. Am. Statist. Assoc.* **86**, 328–332.
- [8] Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.
- [9] Dawid, A. P. (1979). Conditional independence in statistical theory. *J. R. Statist. Soc. B* **41**, 1–31.
- [10] Eaton, Morris L. (1986). A characterization of spherical distributions. *J. Multiv. Anal.* **20**, 272–276.
- [11] Eaton, Morris L. and Tyler, D. E. (1994). The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis. *J. Mult. Anal.* **34**, 439–446.
- [12] Diaconis, Persi and Freedman, David (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815.
- [13] Hall, Peter and Li, Ker-Chau (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* **21**, 867–889.
- [14] Li, Ker-Chau (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* **86**, 316–342.



- [15] Li, Ker-Chau (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Assoc.* **87**, 1025–1039.
- [16] Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: Wiley.

# A reduction paradigm for multivariate laws

Francesca Chiaromonte

*International Institute for Applied Systems Analysis, Austria.*

*Abstract:* A *reduction paradigm* is a theoretical framework which provides a definition of structure for multivariate laws, and allows to simplify their representation and statistical analysis. The main idea is to decompose a law as the superposition of a *structural term* and a *noise*, so that the latter can be neglected *without loss of information on the structure*. When the structural term is supported by a lower-dimensional affine subspace, an *exhaustive dimension reduction* is achieved. We describe the reduction paradigm that results from selecting white noises, and convolution as superposition mechanism.

*Key words:* Multivariate structure, multivariate dimension-reduction, multivariate graphics.

AMS subject classification: 62A99, 62H05.

## 1 Introduction

A  $k$ -variate law is a complex object whose structure embodies both marginal and joint features. All those features can be translated, to some extent, into geometric characterizations of an iid sample from the law, meant as a cloud of points in  $\mathbb{R}^k$ . Dimension does not affect the analysis of marginal features, but as  $k$  increases it becomes progressively harder to conceive and articulate the joint ones. For example, how does one conceive and articulate the interdependencies among, say, 10 or 100 coordinate components? One is often forced to neglect high-order interactions, and/or to assume hierarchies among them <sup>1</sup>. At the same time, for  $k > 3$ , the data cannot

---

<sup>1</sup>Conditional independence (see A.P. Dawid, 1979) provides a key to articulate interdependencies; a very interesting representation of them through *conditional independence*

be visualized as a whole; while graphical tools can still be used to investigate low-dimensional marginals, a *direct* graphical investigation of the joint features is impossible.

Producing inferences in high-dimensional settings can then become complicated and challenging. A large variety of inference methods is available once strong assumptions on the nature of the law are imposed; that is, once a model for the law is chosen (see, among others, M.L. Eaton, 1983, R.J. Muirhead, 1982, and G.A.F. Seber, 1984). But the intuition based on graphical preliminary exploration that should precede the utilization of model-based methods is impaired by the conceptual and practical difficulties mentioned above.

These considerations, among others, justify the quest for simplified representations of multivariate laws, especially ones allowing a *reduction in dimension*. Simplified representations are often developed targeting some (more or less restricted) features of interest. *Exhaustiveness* becomes then an issue; once a *target* has been chosen, the information concerning it ought to be preserved by simplification. More generally, it ought to be clear in what relation the proposed simplified representation is to the target. If exhaustiveness is not always guaranteed, it should be possible to state under what assumptions on the nature of the law it is, and/or to establish to what extent the target is preserved (with or without assumptions).

These issues are very relevant in practice; the last thirty years have witnessed the development of a large number of graphical exploration procedures for high-dimensional data sets. Think for example of Principal Component Analysis, Factor Analysis (see G.A.F. Seber, 1984, and references therein), Projection Pursuit (H.J. Friedman and J.W. Tuckey, 1974, H.J. Friedman, 1987, and D. Cook, A. Buja, J. Cabrera and C. Hurley, 1995), or Grand Tours (D. Asimov, 1985, and A. Buja and D. Asimov, 1986). The theoretical rationale underlying any of these procedures can be interpreted as a simplified representation of the multivariate law from which the data are drawn; targets range anywhere from “variability”, to “linear interdependence structure” (correlation among the coordinate components), to “non-linear structure” (defined as departure from normality), to “structure” according to some other definition. Correspondingly, many of the critiques to these procedures can be interpreted in terms of choice of targets, and relations between simplified representations and targets. As we proceed, it will become clear that the simplified representation underlying Factor Analysis is the closest in spirit to the one we will propose. In fact, Factor Analysis differs from the other procedures mentioned above by

---

*graphs* is given by J. Whittaker, 1990.

its reference to a *latent factor* entirely embodying the correlation target.

Our focus will not be on techniques to make inference on simplified representations (“population” objects) based on data from a multivariate law, but on the theoretical premises for these techniques; that is, on how to define targets, and how to develop simplified representations guaranteed to embody them exhaustively.

In Sections 1 and 2, we introduce the concept of *reduction paradigm* and provide definitions and some key results. Section 3 concerns dimension reduction. We conclude with a brief summary and some remarks on inference in Section 4. More details can be found in F. Chiaromonte, 1996.

## 2 The reduction paradigm

Our analysis will be conducted at the level of laws on  $\mathbb{R}^k$ , and we will not distinguish among random vectors with the same distribution. The main idea behind a reduction paradigm is to decompose a law  $L$  on  $\mathbb{R}^k$  into two terms, one of which does not contribute to the structure (the target) and can therefore be neglected. In other words, the aim is to represent a law as the superposition of a *structural term* and a *noise*, or *no-structure term*. Hence, the specification of a reduction paradigm relies upon

- a definition of absence of structure; that is, a choice of noises
- a choice of superposition mechanism

which, conversely, determine a definition of structure. We have selected *white noises*  $N_k(0, \beta I_k)$ ,  $\beta \in \mathbb{R}_+^1$ , and *convolution*. Hence, we write

$$L = \Lambda_\beta(L) * N_k(0, \beta I_k) \quad (1)$$

or, in terms of characteristic functions

$$\phi_L(u) = \phi_{\Lambda_\beta(L)}(u) e^{-\frac{\beta}{2} \|u\|^2} \quad ::, :: u \in \mathbb{R}^k \quad (2)$$

This is by no means the only possibility, but it is in line with much of the statistical tradition and thus constitutes a very natural first step. In fact, it expresses a situation in which an independent normal error is additively superimposed to the object of interest. One can envision reproducing the whole analysis we are about to develop with different noises and/or superposition mechanisms, though. As far as noises are concerned, one could take, for example, uniforms on hyper-spheres of radius  $\rho \in \mathbb{R}_+^1$ , or normals with independent components  $N_k(0, \text{Diag}(\sigma_i))$ ,  $\sigma \in \mathbb{R}_+^k$ . In the first case one maintains the weakly spherical nature of white noises and loses independence of the coordinate components, while in the second case one loses weak sphericity and maintains independence. Regarding superposition, one could explore, for example, multiplicative (instead of additive) schemes.

Before proceeding let us remark that the reduction paradigm we have selected, as well as any other conceivable one, while certainly constituting a *model for decomposing a law*, does not require strong assumptions on the nature of the law itself. A reduction paradigm can be applied without fixing at the outset a *model for the law*; that is, without assuming at the outset that the law belongs to a given (and possibly finally parameterized) class. Furthermore, our reduction paradigm corresponds to the inverse problem for heat-type diffusion of probability measures (for an easy introduction, see G.M. Wing, 1991, and A. Friedman and W. Littman, 1994). Paradigms resulting from a different choice of noises would correspond to inverse problems for processes with different kernels.

Indexing the structural term by  $\beta$  serves to stress the fact that the decomposition in (1) and (2) is not unique, unless it holds only with  $\beta = 0$ , and therefore  $\Lambda_\beta(L) = L$  itself. The set

$$\mathcal{B}(L) = \left\{ \beta \in \mathbb{R}_+^1 \text{ s.t. } \phi_L(\cdot) e^{\frac{\beta}{2} \|\cdot\|^2} \text{ is a ch. fct.} \right\} \subseteq \mathbb{R}_+^1$$

expresses the range of possible decompositions. It is always non-empty, as it must contain 0, and is easily shown to be  $\mathcal{B}(L) = [0, \beta_o(L)]$ , where  $\beta_o(L) = \sup \mathcal{B}(L) = \max \mathcal{B}(L)$ . We call the corresponding structural terms

$$\Lambda_\beta(L) \leftrightarrow \phi_L(\cdot) e^{\frac{\beta}{2} \|\cdot\|^2}, \quad \beta \in \mathcal{B}(L)$$

*sources*,  $\beta_o(L)$  *reduction coefficient*, and  $\Lambda_o(L) \leftrightarrow \phi_L(\cdot) e^{\frac{\beta_o(L)}{2} \|\cdot\|^2}$  *primary source* of  $L$ . Notice that reduction coefficient and primary source are unique by construction. If  $\beta_o(L) = 0$ , so that the only (and thus primary) source of  $L$  is  $L$  itself, we say that the law is *irreducible*. We call it *reducible* otherwise.

All sources share the structure of  $L$ , and can be equivalently taken as exhaustive “simplified” representations of the law. The primary source is the one in which no error is superimposed to the structure; that is, the one in which we have pushed simplification as far as possible. Hence, we will select  $\Lambda_o(L)$  as simplified representation of  $L$ , and write

$$L = \Lambda_o(L) * N_k(0, \beta_o(L) I_k)$$

We can fix ideas using the normal case as an example. Here and in the following,  $P_{(\cdot)}^2$  indicates the orthogonal projection operator onto the argument subspace<sup>2</sup>, with respect to the standard inner product on  $\mathbb{R}^k$ . Let

$$L = N_k \left( \mu, \sum_{j=1}^p \eta_j P_{V_j} + \eta P_V \right)$$

---

<sup>2</sup>The reference, throughout our discussion, is to *linear* subspaces, and affine subspaces obtained by translating them.

where  $\sum_{j=1}^p \eta_j P_{V_j} + \eta P_V$  is the spectral decomposition of the covariance with (distinct) eigenvalues  $\eta_1, \dots, \eta_p, \eta$  in decreasing order, and corresponding eigenspaces  $V_1, \dots, V_p, V$ . It is easy to show that

$$\phi_L(u) e^{\frac{\theta}{2} \|u\|^2} = \exp \left\{ i u' \mu - \frac{1}{2} u' \left( \sum_{j=1}^p (\eta_j - \beta) P_{V_j} + (\eta - \beta) P_V \right) u \right\}$$

is a characteristic function if and only if  $\sum_{j=1}^p (\eta_j - \beta) P_{V_j} + (\eta - \beta) P_V$  is non-negative definite; that is, if and only if  $\beta \leq \eta$ . Hence,  $\beta_o(L) = \eta$  and correspondingly

$$\Lambda_o(L) = N_k \left( \mu, \sum_{j=1}^p (\eta_j - \eta) P_{V_j} \right)$$

It is then clear that a normal is irreducible if and only if the smallest eigenvalue of its covariance is 0; the irreducible  $k$ -variate normals are all and only the ones supported by lower dimensional affine subspaces, and they constitute the primary sources of non-singular normals.

Primary sources are irreducible by construction. The class of all irreducible laws on  $\mathbb{R}^k$  represents the repertoire of possible structures. The following proposition provides a *sufficient condition for irreducibility*, thereby characterizing part of such repertoire.

**Proposition 1** *If there exists a measurable set  $B \subseteq \mathbb{R}^k$  such that  $\text{Leb}(B) > 0$ , but  $L(B) = 0$ , then  $L$  is irreducible.*

**Proof:** Suppose  $\beta_o(L) > 0$ . Then, for any choice of  $v \in \mathbb{R}^k$ ,  $N_k(v, \beta_o(L) I_k)$  is mutually absolutely continuous with respect to  $\text{Leb}$ . So  $\text{Leb}(B) > 0$  implies

$$N_k(v, \beta_o(L) I_k)(B) > 0, \quad \forall v \in \mathbb{R}^k$$

and thus

$$L(B) = \int_{\mathbb{R}^k} N_k(v, \beta_o(L) I_k)(B) \Lambda_o(L)(dv) > 0$$

contradicting our assumption. We can conclude that  $\beta_o(L) = 0$ , and therefore that  $L$  is irreducible.  $\square$

Since we have selected white noises as no-structure terms, reducible laws must be mutually absolutely continuous with respect to the Lebesgue measure, because they “contain” a term that is. As a consequence, all laws having “thick” holes with respect to the Lebesgue measure are irreducible in  $\mathbb{R}^k$ . In particular, laws whose affine support  $\text{As}(L)$  has dimension  $< k$  are irreducible in  $\mathbb{R}^k$ ; we saw an instance of this with irreducible normals. So are laws whose closed support  $\text{Cs}(L)$  is bounded, regardless of whether

the latter is full-dimensional or embedded in a subspace or affine subspace of dimension  $< k$ .

Notice that existence of an everywhere positive density is *not* enough to guarantee reducibility; again because of our choice of no-structure terms, reducible laws' densities must have "thick enough" tails. It is easy to show that a law with an everywhere positive density whose tails vanish too fast, at least along some directions, will still be irreducible (see F. Chiaromonte, 1996).

### 3 Some affine actions, and marginalizations

We will now explore the effects on reduction of some affine actions and of marginalizations.

**Proposition 2** *Let  $\mathcal{T}_{v,r,R}[L]$  be the law of  $rRX - v$ , where  $X \in \mathbb{R}^k$  is any random vector distributed according to  $L$ ,  $v \in \mathbb{R}^k$ ,  $r \in \mathbb{R}^1$ , and  $R$  is a rotation of  $\mathbb{R}^k$ . Then  $\beta_o(\mathcal{T}_{v,r,R}[L]) = \beta_o(L)$  and  $\Lambda_o(\mathcal{T}_{v,r,R}[L]) = \mathcal{T}_{v,r,R}[\Lambda_o(L)]$ .*

**Proof:** For  $r = 0$ , the transformation yields a point-mass at  $-v$ , and the statement is trivially true. Otherwise, using characteristic functions, one has

$$\begin{aligned} \phi_{\mathcal{T}_{v,r,R}[L]}(u) &= e^{-iu'v} \phi_L(rR'u) \\ &= e^{-iu'v} \phi_{\Lambda_o(L)}(rR'u) e^{-\frac{\beta_o(L)}{2} \|rR'u\|^2} \\ &= \phi_{\mathcal{T}_{v,r,R}[\Lambda_o(L)]}(u) e^{-\frac{r^2 \beta_o(L)}{2} \|u\|^2} \end{aligned}$$

so  $\beta_o(\mathcal{T}_{v,r,R}[L]) \geq r^2 \beta_o(L)$ , and  $\mathcal{T}_{v,r,R}[\Lambda_o(L)]$  is a source of  $\mathcal{T}_{v,r,R}[L]$ . But for  $r \neq 0$  our transformation is invertible:  $\mathcal{T}_{v,r,R}^{-1}[\cdot] = \mathcal{T}_{-v,1/r,R'}[\cdot]$ . Hence

$$\begin{aligned} \phi_L(u) &= \phi_{\mathcal{T}_{-v,1/r,R'} \mathcal{T}_{v,r,R}[L]}(u) \\ &= \phi_{\mathcal{T}_{-v,1/r,R'}[\Lambda_o(\mathcal{T}_{v,r,R}[L])]}(u) e^{-\frac{(1/r)^2 \beta_o(\mathcal{T}_{v,r,R}[L])}{2} \|u\|^2} \end{aligned}$$

and  $\beta_o(L) \geq (1/r)^2 \beta_o(\mathcal{T}_{v,r,R}[L])$ . We can conclude that  $\beta_o(\mathcal{T}_{v,r,R}[L]) = r^2 \beta_o(L)$ , and therefore that  $\mathcal{T}_{v,r,R}[\Lambda_o(L)]$  is indeed the primary source of  $\mathcal{T}_{v,r,R}[L]$ .  $\square$

The reduction coefficient is not affected by rotations and translations, and is multiplied by the square of a rescaling factor. Thus, rescalings, rotations and translations of  $L$  result into corresponding rescalings, rotations and translations of the primary source. In the following, we will use

interchangeably the terms *marginalization* and *projection*. Besides the intuitive correspondence, “invariance” under rotations makes this rigorous; the choice of orthonormal basis does not matter.

In our discussion so far, we have considered the reduction of a law  $L$  on  $\mathbb{R}^k$  in  $\mathbb{R}^k$ . The reference to the space is important; laws on  $\mathbb{R}^k$  that are entirely concentrated on some subspace can also be meant as laws on such subspace, and reducing them *within the subspace* can produce a different set of sources, a different reduction coefficient and a different primary source. Laws that are entirely concentrated on a subspace of dimension  $< k$  are irreducible in  $\mathbb{R}^k$ , but they might still be reducible within the subspace.

The noises within a given subspace  $S \subseteq \mathbb{R}^k$  are represented by  $N_k(0, \beta P_S)$ ,  $\beta \in \mathbb{R}_+^1$ . Notation-wise, when considering the reduction of a law  $\tilde{L}$  (entirely concentrated on  $S$ ) within  $S$ , we will write  $\beta_o(\tilde{L}, S)$ ,  $\Lambda_o(\tilde{L}, S)$ , etc.

**Proposition 3** *Let  $\mathcal{M}_S[L]$  be the law of  $P_S X$ , where  $X \in \mathbb{R}^k$  is any random vector distributed according to  $L$ , and  $S \subseteq \mathbb{R}^k$  is a non-degenerate subspace. Then  $\beta_o(\mathcal{M}_S[L], S) \geq \beta_o(L)$  and*

$$\Lambda_o(\mathcal{M}_S[L], S) * N_k(0, \alpha P_S) = \mathcal{M}_S[\Lambda_o(L)]$$

where  $\alpha = \beta_o(\mathcal{M}_S[L], S) - \beta_o(L)$ . In particular, if  $\text{Cs}(\Lambda_o(L))$  is bounded,  $\beta_o(\mathcal{M}_S[L], S) = \beta_o(L)$  and  $\Lambda_o(\mathcal{M}_S[L], S) = \mathcal{M}_S[\Lambda_o(L)]$ .

**Proof:** Using characteristic functions, one has

$$\begin{aligned} \phi_{\mathcal{M}_S[L]}(u) &= \phi_L(P_S u) \\ &= \phi_{\Lambda_o(L)}(P_S u) e^{-\frac{\beta_o(L)}{2} \|P_S u\|^2} \\ &= \phi_{\mathcal{M}_S[\Lambda_o(L)]}(u) e^{-\frac{\beta_o(L)}{2} \|P_S u\|^2} \end{aligned}$$

so  $\beta_o(\mathcal{M}_S[L], S) \geq \beta_o(L)$ , and  $\mathcal{M}_S[\Lambda_o(L)]$  is a source of  $\mathcal{M}_S[L]$  within  $S$ . Equating the right hand side above with the right hand side of

$$\phi_{\mathcal{M}_S[L]}(u) = \phi_{\Lambda_o(\mathcal{M}_S[L], S)}(u) e^{-\frac{\beta_o(\mathcal{M}_S[L], S)}{2} \|P_S u\|^2}$$

we obtain

$$\phi_{\Lambda_o(\mathcal{M}_S[L], S)}(u) e^{-\frac{\alpha}{2} \|P_S u\|^2} = \phi_{\mathcal{M}_S[\Lambda_o(L)]}(u)$$

where  $\alpha = \beta_o(\mathcal{M}_S[L], S) - \beta_o(L)$ ; that is

$$\Lambda_o(\mathcal{M}_S[L], S) * N_k(0, \alpha P_S) = \mathcal{M}_S[\Lambda_o(L)]$$

Now, assume that  $\text{Cs}(\Lambda_o(L))$  is bounded. We need to show that this implies  $\alpha = 0$ . Suppose  $\alpha > 0$ . Then, because of the normal term

$$\text{Cs}(\mathcal{M}_S[\Lambda_o(L)]) = \text{Cs}(\Lambda_o(\mathcal{M}_S[L], S) * N_k(0, \alpha P_S)) = S$$



But if  $\text{Cs}(\mathcal{M}_S[\Lambda_o(L)])$  is unbounded  $\text{Cs}(\Lambda_o(L))$  must be unbounded, too, contradicting our assumption. We can conclude that  $\beta_o(\mathcal{M}_S[L], S) = \beta_o(L)$ , and therefore that  $\Lambda_o(\mathcal{M}_S[L], S) = \mathcal{M}_S[\Lambda_o(L)]$ .  $\square$

The reduction coefficient (within  $S$ ) of the marginal of  $L$ , must be greater than or equal to  $\beta_o(L)$ . Correspondingly, the marginal of  $\Lambda_o(L)$  is a source (within  $S$ ) of the marginal of  $L$ , even though not necessarily the primary one. Under the assumption that  $\text{Cs}(\Lambda_o(L))$  is bounded, the reduction coefficients coincide. Thus, the marginal of  $\Lambda_o(L)$  is indeed the primary source (within  $S$ ) of the marginal of  $L$ . In other words, under the boundedness assumption the reduction coefficient is not affected by marginalizations (projections), and therefore marginalizations of  $L$  result into corresponding marginalizations of the primary source.

## 4 The structural subspace, and exhaustive dimension reduction

The affine support of  $\Lambda_o(L)$  represents the smallest affine subspace supporting the structure of  $L$ , as defined by our reduction paradigm. We call the subspace underlying it the *structural subspace* of the law  $S_o(L) = \text{As}(\mathcal{T}_v[\Lambda_o(L)])$ , where  $v$  is any element of  $\text{Cs}(\Lambda_o(L))$ , and  $\mathcal{T}_v$  stands for  $\mathcal{T}_{v,1,I_k}$ . Correspondingly, we call  $d_o(L) = \dim(S_o(L))$  the *structural dimension*. Whenever  $d_o(L) < k$ , our (exhaustive) simplified representation of  $L$  implies a drop in dimension.

This allows us to define an *exhaustive dimension reduction*. Let us see how. Suppose we know  $v \in \text{Cs}(\Lambda_o(L))$ . Then, the exercise of identifying  $\Lambda_o(L)$  is equivalent to that of identifying  $\Lambda_o(\mathcal{T}_v[L])$ . In fact, by Proposition 2

$$\Lambda_o(L) = \mathcal{T}_{-v}\mathcal{T}_v[\Lambda_o(L)] = \mathcal{T}_{-v}[\Lambda_o(\mathcal{T}_v[L])]$$

Now, suppose  $S_o(L)$  is known, too. Then, we can marginalize  $\mathcal{T}_v[L]$  to  $S_o(L)$  preserving all the information relative to the structure, as defined by our reduction paradigm. In fact, again by Proposition 2

$$S_o(L) = \text{As}(\mathcal{T}_v[\Lambda_o(L)]) = \text{As}(\Lambda_o(\mathcal{T}_v[L]))$$

so that indeed  $\Lambda_o(\mathcal{T}_v[L])$  is supported by the structural subspace<sup>3</sup>, and

$$\Lambda_o(\mathcal{T}_v[L]) = \mathcal{M}_{S_o(L)}[\Lambda_o(\mathcal{T}_v[L])]$$

---

<sup>3</sup>Notice that, by Proposition 2, the structural subspace is invariant under translations of  $L : S_o(\mathcal{T}_v[L]) = S_o(L)$ . When translating by an element of  $\text{Cs}(\Lambda_o(L))$  we obtain a law which is actually supported by the subspace itself, instead of an affine subspace parallel to it.

But then, by Proposition 3

$$\Lambda_o(\mathcal{T}_v[L]) = \Lambda_o(\mathcal{M}_{S_o(L)} \mathcal{T}_v[L], S_o(L)) * N_k(0, \alpha P_{S_o(L)})$$

$\Lambda_o(\mathcal{T}_v[L])$  is a source for  $\mathcal{M}_{S_o(L)} \mathcal{T}_v[L]$  within  $S_o(L)$ . Furthermore, if we can assume  $\text{Cs}(\Lambda_o(L))$ , and therefore of  $\text{Cs}(\Lambda_o(\mathcal{T}_v[L]))$ , to be bounded

$$\Lambda_o(\mathcal{T}_v[L]) = \Lambda_o(\mathcal{M}_{S_o(L)} \mathcal{T}_v[L], S_o(L))$$

that is,  $\Lambda_o(\mathcal{T}_v[L])$  is the primary source of  $\mathcal{M}_{S_o(L)} \mathcal{T}_v[L]$  within  $S_o(L)$ . This gives an even stronger meaning to the exhaustiveness of our marginalization; not only no structural information is lost marginalizing  $\mathcal{T}_v[L]$ , but the exercise of identifying  $\Lambda_o(\mathcal{T}_v[L])$  (to be performed in  $k$  dimensions) would actually correspond to that of identifying  $\Lambda_o(\mathcal{M}_{S_o(L)} \mathcal{T}_v[L], S_o(L))$  (to be performed in –possibly– smaller dimension).

The question becomes then how to identify translation term and structural subspace. Clearly, existence of finite moments of a certain order for  $L$  implies that of the corresponding moments for  $\Lambda_o(L)$ . If  $L$  admits finite first order moments  $E(\Lambda_o(L)) = E(L)$ , and one can take as translation term  $v = E(L) \in \text{Cs}(\Lambda_o(L))$ . Furthermore, if  $L$  admits finite second order moments, structural subspace and structural dimension can be related to the spectral decomposition of the covariance. We will need the following

**Lemma 1** *If  $L$  admits finite second order moments, then  $\text{As}(\mathcal{T}_{E(L)}[L]) = \text{Span}(\text{Cov}(L))$ .*

**Proof:** Consider the orthogonal complement of  $\text{Span}(\text{Cov}(L))$  with respect to the standard inner product, and  $\mathcal{M}_{\text{Span}(\text{Cov}(L))^\perp} \mathcal{T}_{E(L)}[L]$ . Easy calculations give

$$E(\mathcal{M}_{\text{Span}(\text{Cov}(L))^\perp} \mathcal{T}_{E(L)}[L]) = 0, \text{Cov}(\mathcal{M}_{\text{Span}(\text{Cov}(L))^\perp} \mathcal{T}_{E(L)}[L]) = 0$$

Thus,  $\mathcal{M}_{\text{Span}(\text{Cov}(L))^\perp} \mathcal{T}_{E(L)}[L]$  is a point-mass at 0, which implies  $\text{Span}(\text{Cov}(L)) \supseteq \text{As}(\mathcal{T}_{E(L)}[L])$ . On the other hand, using the definition of covariance we have

$$\text{Cov}(L)z = \text{Cov}(\mathcal{T}_{E(L)}[L])z = \int_{\text{Cs}(\mathcal{T}_{E(L)}[L])} vv'z \mathcal{T}_{E(L)}[L](dv)$$

which implies  $\text{Span}(\text{Cov}(L)) \subseteq \text{As}(\mathcal{T}_{E(L)}[L])$ . The statement follows.  $\square$

Now, denoting by  $\text{Ind}(\cdot)$  the indicator function of the argument condition, we have

**Proposition 4** *Suppose  $L$  admits finite second order moments. Let  $\eta(L)$  be the smallest eigenvalue of  $\text{Cov}(L)$ , and  $V(L)$  the corresponding eigenspace. Then  $\beta_o(L) \leq \eta(L)$  and*

$$S_o(L) = V(L)^\perp \oplus \text{Ind}(\eta(L) - \beta_o(L) > 0)V(L)$$

with  $d_o(L) = k - \text{Ind}(\eta(L) - \beta_o(L) = 0)\dim(V(L))$ .

**Proof:** Writing  $\text{Cov}(L) = \sum_{j=1}^p \eta_j(L)P_{V_j(L)} + \eta(L)P_{V(L)}$  one has

$$\begin{aligned} \text{Cov}(\Lambda_o(L)) &= \text{Cov}(L) - \beta_o(L)I_k \\ &= \sum_{j=1}^p (\eta_j(L) - \beta_o(L))P_{V_j(L)} + (\eta(L) - \beta_o(L))P_{V(L)} \end{aligned}$$

But then  $\beta_o(L) \leq \eta(L)$  is implied by non-negative definiteness of  $\text{Cov}(\Lambda_o(L))$ , and the expression for the structural subspace follows from Lemma 1 applied to  $\Lambda_o(L)$ . A drop in dimension occurs if and only if  $\beta_o(L) = \eta(L)$ , and when it occurs  $d_o(L) = k - \dim(V(L))$ , where  $\dim(V(L))$  represents the multiplicity of  $\eta(L)$ .  $\square$

Given the spectral decomposition of  $\text{Cov}(L)$ , the above proposition provides an upper bound for the reduction coefficient and a lower bound for the structural subspace; namely, the smallest eigenvalue of  $\text{Cov}(L)$  and the orthogonal complement of its eigenspace. The spectral decomposition of  $\text{Cov}(L)$  is not enough to identify the structural subspace, though; we still need to know whether the reduction coefficient is strictly smaller than, or equal to, the smallest eigenvalue.

Remember that for a normal law  $\beta_o(L) = \eta$ . Thus, under normality the drop in dimension always occurs, and one has  $S_o(L) = V^\perp$  with  $d_o(L) = k - \dim(V) \leq k - 1$ . It is important to remark that coincidence of  $\beta_o(L)$  with  $\eta(L)$  (and therefore the drop in dimension) is not guaranteed in general. Identifying the reduction coefficient with the smallest eigenvalue of the covariance can actually be very misleading. Take for example a “noisy” uniform on a hyper-cube  $L = Un([- \theta, \theta]^k) * N_k(0, \tau I_k)$ ,  $\theta, \tau \in \mathbb{R}_+^1 \setminus \{0\}$ . For such a law one has  $\beta_o(L) = \tau < \frac{\theta^2}{2} + \tau = \eta(L)$  and  $S_o(L) = \mathbb{R}^k \supset \{0\} = V(L)^\perp$ , as the multiplicity of  $\frac{\theta^2}{2} + \tau$  is  $k$ .

## 5 A brief summary with some remarks on inference

The ultimate aim within the framework defined by a reduction paradigm is that of making inference about the (unobservable)  $\Lambda_o(L)$ , which constitutes

our simplified and yet exhaustive representation of the original law. As we have seen, if we can assume that  $L$  is normal,  $\beta_o(L) = \eta$  and

$$\Lambda_o(L) = N_k \left( \mu, \sum_{j=1}^p (\eta_j - \eta) P_{V_j} \right)$$

which is entirely identified through the mean vector and the spectral decomposition of the covariance. Hence, if the data are consistent with normality, we could estimate the primary source based on estimates of those. Also, if the data can be transformed to approximate normality, the primary source could be estimated on the transformed scale. What can we do when the data contradicts normality on the original scale, and fails to approximate it also after applying normalizing transformations?

An intermediate aim is constituted by estimating a  $v \in \text{Cs}(\Lambda_o(L))$  and  $S_o(L)$ . Besides the intrinsic interest, if indeed our simplified representation implied a drop in dimension, having such estimates would allow us to perform an exhaustive dimension reduction.

Given the results in the previous sections, we are clearly at an advantage if we are willing to assume boundedness of  $\text{Cs}(\Lambda_o(L))$ . Since the latter implies existence and finiteness for all the moments of  $L$ , we would have  $E(L) \in \text{Cs}(\Lambda_o(L))$  and (Proposition 4)

$$S_o(L) = V(L)^\perp \oplus \text{Ind}(\eta(L) - \beta_o(L) > 0)V(L)$$

Furthermore, we could restrict inference on the reduction coefficient to any arbitrarily small non-degenerate subspace. In fact, by Proposition 3  $\beta_o(L) = \beta_o(\mathcal{M}_t[L], t)$ , where  $t$  is any line in  $\mathbb{R}^k$ . Thus, we could take  $\hat{E}(L)$  as translation term, and produce an estimate of the structural subspace based on  $\hat{\eta}(L)$ ,  $\hat{V}(L)$  and  $\hat{\beta}_o(\mathcal{M}_t[L], t)$ . Methods to estimate  $E(L)$ , and, less trivially,  $\eta(L)$  and  $V(L)$ , exist in the literature and are not affected by how large  $k$  is (see M.L. Eaton and D. Tyler, 1994, and E. Bura, 1996).

As a matter of fact, in order to produce an estimate of the structural subspace we would only have to assess, selecting for example  $t \subseteq \hat{V}$ , whether  $\beta_o(\mathcal{M}_t[L], t)$  is strictly smaller than, or coincides with  $\text{var}(\mathcal{M}_t[L]) = \eta(L)$ . This, in turn, is equivalent to assessing whether  $\mathcal{M}_t[L]$  is a 1-dimensional normal.

Under the assumption that  $\text{Cs}(\Lambda_o(L))$  is bounded, we also have that

$$\Lambda_o(L) = \mathcal{T}_{-v}[\Lambda_o(\mathcal{M}_{S_o(L)}\mathcal{T}_v[L], S_o(L))]$$

Hence, we could center the data cloud translating it by  $\hat{E}(L)$ , and restrict any further analysis to the projection of the centered cloud onto  $\hat{S}_o(L)$ ;

all the structural features (except for location, which is captured by  $\hat{E}(L)$ ) would be preserved. If indeed  $\hat{d}_o(L) = \dim(\hat{S}_o(L)) < k$ , we would have achieved an exhaustive dimension reduction.

## References

- [1] Asimov D. (1985). The grand tour: a tool for viewing multidimensional data. *SIAM J. Scient. Statist. Comput.* **6**, 128–143.
- [2] Buja A., and Asimov D. (1986). Grand tour methods: an outline. *Computing Science and Statistics* **17**, 63–67.
- [3] Bura E. (1996). Dimension reduction via inverse regression. Ph.D. Dissertation, School of Statistics, University of Minnesota.
- [4] Chiaromonte F. (1996). A reduction paradigm for multivariate laws. Ph.D. Dissertation, School of Statistics, University of Minnesota.
- [5] Cook D., Buja A., Cabrera J., and Hurley C. (1995). Grand tour and projection pursuit. *J. Comput. Graph. Statist.* **4** 155–172.
- [6] Dawid A.P. (1979). Conditional independence in statistical theory. *J. R. Statist. Soc.B* **41**, 1–31.
- [7] Eaton M.L. (1983). *Multivariate Statistics*. New York: Wiley.
- [8] Eaton M.L., and Tyler D. (1994). The asymptotic distribution of singular values, with applications to canonical correlations and correspondence analysis. *J. Mult. Anal.* **50**.
- [9] Friedman A., and Littman W. (1994). *Industrial Mathematics: a Course in Solving Real World Problems*. Philadelphia: SIAM.
- [10] Friedman H.J. (1987). Exploratory projection pursuit. *J. Am. Statist. Assoc.* **82**, 249–266.
- [11] Friedman H.J., and Tuckey J.W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* **23**, 881–889.
- [12] Muirhead R.J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.
- [13] Posse C. (1995). Tools for two-dimensional exploratory projection pursuit. *J. Comput. Graph. Statist.* **4**, 83–100.
- [14] Seber G.A.F. (1984). *Multivariate Observations*. New York: Wiley.
- [15] Whittaker J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.
- [16] Wing G.M. (1991). *A Primer on Integral Equations of the First Kind: the Problem of Deconvolution and Unfolding*. SIAM, Port City Press.

# Exploring data sets using partial residual plots based on robust fits

**Joseph W. McKean**

*Western Michigan University, USA*

**Simon J. Sheather**

*University of New South Wales, Australia*

*Abstract:* Partial residual plots are one of the most useful graphical procedures in the exploratory fitting of data sets. They are frequently used in the identification of unknown functions,  $g$ , of predictor variables. Traditionally these plots have been based on least squares (LS) fitting. It is well known that LS estimates are sensitive to outlying observations. The examples and sensitivity study discussed in this paper show that this vulnerability to outliers carries over to the LS based partial residual plots. A few outliers in the data set can distort the LS partial residual plot making the identification of  $g$  impossible. Furthermore, if  $g$  is non-linear, good data points may act as outliers and cause distortion in the plot. Partial residual plots based on highly efficient robust estimates are presented. In the simulated data sets explored in this paper, the robust based partial residual plots are insensitive to the outlying observations leading to a much easier identification of the unknown functions than their LS counterparts. In the sensitivity study presented, these robust based partial residual plots do not become distorted in the presence of outliers but they maintain their focus, enabling the identification of  $g$ .

*Key words:* Linear model, M-estimates, outlier, regression diagnostics, R-estimates, rank based methods.

AMS subject classification: 62J20, 62G35.

## 1 Introduction

Partial residual plots are one of the most useful graphical procedures in the

exploratory fitting of data sets. These plots are quite simple. Consider a model of the form  $y_i = \alpha + \beta'_1 \mathbf{x}_{i1} + g(x_{i2}) + \epsilon_i$ , where the function  $g(x)$  is unknown. Then (the first-order) partial residuals are the residuals of the fit of the misspecified model  $y_i = \alpha + \beta'_1 \mathbf{x}_{i1} + \beta_2 x_{i2} + e_i$  added to the fitted part  $\hat{\beta}_2 x_{i2}$ . The plot consists of these partial residuals plotted versus  $x_{i2}$ . This plot is often informative in the identification of the unknown function  $g(x)$ .

Partial residual plots were proposed by Ezekiel (1924) and have been discussed by numerous authors. Larsen and McCleary (1972) gave the name partial residual plot to this procedure. Mallows (1986) extended these first-order plots to higher orders, the so-called augmented partial residual plots; Mansfield and Conerly (1987) considered informative algebraic representations of partial residuals; Cook (1993) obtained further theoretical underpinnings of these plots and proposed an extended class, the CERES plots; and Berk and Booth (1995) compare partial residual plots with several other diagnostic plots in a series of interesting examples. Based on work such as this, partial residual plots have become an important tool in data exploration.

Most of the discussion of partial residual plots is based on the traditional least squares (LS) fitting of models. Partial residuals, though, are simply residuals added to the fit of the misspecified part. Hence, fits other than LS can be considered. McKean and Sheather (1997) developed properties of partial residuals based on robust fitting. They showed that the expected behavior of the resulting robust partial residual plots was similar to that of the LS partial residual plots. Furthermore, they showed that the robust partial residual plots were not as sensitive to outliers as the LS based plots.

To determine which robust estimates to use, note that the function  $g(x)$  is often a nonlinear function. Hence the employed fitting criteria should be able to detect and fit curvature. We have selected highly efficient M and R estimates as the basis of our fitting criteria. These estimates and their residuals have been shown to behave similar to their LS counterparts in detecting and fitting curvature on good data, while being much less sensitive to LS procedures on data containing outliers in the  $Y$ -space; see McKean, Sheather and Hettmansperger (1990, 1993, and 1994). These fitting criteria are based on minimizing convex functions; hence, the consistency theory developed by Cook (1993) for LS partial residual plots extends to these robust partial residual plots. Also these highly efficient robust fitting criteria are computationally fast and available.

In this paper, we explore several data sets using robust based partial residual plots. In many of these data sets, outliers distort the LS based plots to the point where the identification of the unknown function  $g(x)$  is

impossible. In one of the data sets, due to the nonlinearity of the function  $g$ , good data acted as outliers and distorted the LS based partial residual plot. The robust based partial residual plots, though, are not sensitive to the effect of the outliers. These plots clearly identify the unknown function  $g$ . The sensitivity study in Section 5 shows the distortion of the LS based partial residual plot in a sequential fashion as a few points become increasingly outlying. The robust based partial residual plots, however, retain their “focus” under the increasing influence of the outliers.

## 2 Notation

This paper considers partial residual plots based on robust estimates. As discussed in Section 3, these plots are often used to graphically determine unknown functions of predictors. These functions are often nonlinear so fitting procedures which can detect curvature are of interest. Studies by Cook, Hawkins and Weisberg (1992) and McKean, Sheather and Hettmansperger (1993, 1994) have shown that high breakdown and bounded influence estimates have problems in detecting and fitting curvature, while highly efficient robust estimates are capable of detecting and fitting curvature. Hence, in this article we will focus on highly efficient robust estimates. To keep things simple, we have chosen the Huber M estimate and the Wilcoxon R estimate. Both of these estimates are widely available. But clearly other robust estimates, (other  $\psi$ -functions and other score functions), can be used and will produce similar results. Similar to LS-estimates, though, the Huber and Wilcoxon estimates are highly sensitive to outliers in the  $\mathbf{x}$ -space. This should be considered in exploring any data set prone to outliers in factor space. McKean, Naranjo and Sheather (1996a, 1996b) discuss diagnostic procedures that measure the overall difference between highly efficient and high breakdown robust estimates and determine cases where the fits differ.

Consider the linear regression model  $y_i = \alpha + \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$ ,  $i = 1, \dots, n$  where  $\mathbf{x}_i'$  is the  $i$ th row of the  $n \times p$  centered matrix  $\mathbf{X}$  of explanatory variables defined here. The least squares estimates  $\hat{\alpha}$  and  $\hat{\boldsymbol{\beta}}_{LS}$  minimize the dispersion

$$D_{LS}(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \alpha - \mathbf{x}_i' \boldsymbol{\beta})^2. \quad (1)$$

Let  $\sigma^2$  be the common error variance. Under regularity conditions, the asymptotic distribution of the LS estimates is given by

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_{LS} \end{pmatrix} \sim N \left( \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix}, \sigma^2 \begin{bmatrix} 1/n & \mathbf{0}' \\ \mathbf{0} & (\mathbf{X}'\mathbf{X})^{-1} \end{bmatrix} \right). \quad (2)$$



The regular (Wilcoxon) R-estimate  $\hat{\beta}_W$  minimizes the dispersion

$$D_W(\beta) = \sum_{i=1}^n a(R(y_i - \mathbf{x}'_i \beta))(y_i - \mathbf{x}'_i \beta), \quad (3)$$

where  $R(y_j - \mathbf{x}'_j \beta)$  is the rank of  $y_j - \mathbf{x}'_j \beta$  among  $y_1 - \mathbf{x}'_1 \beta, \dots, y_n - \mathbf{x}'_n \beta$  and the scores  $a(i)$  are generated by the linear function

$$\varphi(u) = \sqrt{12} \left( u - \frac{1}{2} \right), \quad (4)$$

as  $a(i) = \varphi(i/(n+1))$ . Although we will be using Wilcoxon scores throughout this paper, the  $\varphi$  notation will be useful. The function (3) is a convex function of  $\beta$  and Gauss-Newton type algorithms suffice for the minimization; see Kapenga, McKean and Vidmar (1988). Note that (3) is invariant with respect to an intercept term. We shall estimate  $\alpha$  by the median of the Wilcoxon residuals, i.e.,

$$\hat{\alpha}_W = \text{med}(y_i - \mathbf{x}'_i \hat{\beta}_W). \quad (5)$$

Estimating the intercept in this way, avoids unnecessary assumptions such as symmetric error distributions; see Hettmansperger, McKean and Sheather (1997). Our aim is to make as few assumptions as possible when concerned with data exploration.

Under regularity conditions, the Wilcoxon estimates have asymptotic distribution

$$\begin{pmatrix} \hat{\alpha}_W \\ \hat{\beta}_W \end{pmatrix} \sim N \left( \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \begin{bmatrix} \tau_s^2/n & \mathbf{0}' \\ \mathbf{0} & \tau^2(X'X)^{-1} \end{bmatrix} \right), \quad (6)$$

where  $\tau^{-1} = \sqrt{12} \int f^2(t) dt$ , (Jaekel, 1972),  $\tau_s = 1/(2f(0))$ , and  $f$  is the error density. Consistent estimates of  $\tau$  and  $\tau_s$  are presented in Koul, Sievers and McKean (1987) and McKean and Schrader (1983), respectively.

Our third estimate will be Huber's M-estimate  $\hat{\beta}_M$  which minimizes the dispersion function

$$D_M(\beta) = \sum_{i=1}^n \rho((y_i - \alpha - \mathbf{x}'_i \beta)/\sigma_0), \quad (7)$$

where  $\sigma_0$  is a scale parameter and  $\rho$  is given by

$$\rho(x) = \begin{cases} x^2/2 & \text{if } |x| < h, \\ |x|h - h^2/2 & \text{otherwise.} \end{cases} \quad (8)$$

The bend, the parameter  $h$ , must be set. In this paper we will take  $h = 1.345$ , the default setting used in *Splus*; see Becker, Chambers and Wilks (1988). Under regularity conditions  $\hat{\beta}_M$  has an asymptotic normal distribution with asymptotic variance  $\kappa^2(\mathcal{X}'\mathcal{X})^{-1}$ , where  $\mathcal{X} = [\mathbf{1}_n : \mathbf{X}]$  and

$$\kappa^2 = \frac{\sigma^2 E[\psi^2(\epsilon_i/\sigma_0)]}{(E[\psi'(\epsilon_i/\sigma_0)])^2},$$

where  $\psi(t) = \rho'(t)$ ; see Huber (1981). The constant of proportionality  $\kappa^2$  can be estimated by the usual moment estimators.

### 3 Partial residual plots

We will be concerned with models of the form

$$y_i = \alpha + \beta'_1 \mathbf{x}_{1i} + g(\mathbf{x}_{2i}) + \epsilon_i, \quad (9)$$

where  $\mathbf{x}_{1i}$  and  $\mathbf{x}_{2i}$  are  $p \times 1$  and  $q \times 1$  vectors of regression coefficients,  $\mathbf{x}'_{1i}$  is the  $i$ th row of the  $n \times p$  matrix  $\mathbf{X}_1$ , and  $g(x)$  is an unknown function. The goal is to try to determine the function  $g$  as best as possible using simple graphic techniques. The partial residual plot described below is an attempt to obtain this goal. A recent overview can be found in the paper by Cook (1993).

The description of the partial residual plot is the same regardless of what criteria is used to fit a model, so we will describe it generically by dropping the subscripts  $LS$ ,  $W$  and  $M$  for the fitting criteria. Hence, let  $\hat{\beta}$  denote an estimate of the parameter  $\beta$  in a model. We will use the subscripts when distinctions are necessary.

In this article, we will only be looking at cases where the predictor  $x_{i2}$  is univariate. Let  $\mathbf{x}_2 = (x_{12}, \dots, x_{n2})'$ . Since  $g$  is unknown we begin our exploration by fitting a first order model, (at the end of this section we will discuss fitting higher order models). Consider then fitting the model

$$y_i = \alpha + \beta'_1 \mathbf{x}_{1i} + \beta_2 x_{i2} + e_i. \quad (10)$$

Note that, unless  $g(x_{i2}) = \beta_2 x_{i2}$ , model (10) is a misspecified model because model (9) is the correct model. We have indicated this in Model (10) by using  $e_i$  instead of  $\epsilon_i$  for the random error.

Suppose we have fitted the misspecified model (10). Denote the fit by  $\hat{y}_i$  and let  $\hat{e}_i = y_i - \hat{y}_i$  denote the residual. The **partial residuals** are defined by

$$\hat{e}_i^* = \hat{e}_i + \hat{\beta}_2 x_{i2}, \quad (11)$$

that is, the fit of the misspecified part is added back to the residuals. The **partial residual plot** is the plot of  $\hat{e}_i^*$  versus  $x_{i2}$ .

### 3.1 Discussion of partial residual plots

As Cook (1993) noted, since  $\hat{e}_i = y_i - \hat{y}_i$ , we can substitute the right side of equation (9) for  $y_i$  and obtain

$$\hat{e}_i^* = (\alpha - \hat{\alpha}) + (\beta_1 - \hat{\beta}_1)' \mathbf{x}_1 + g(x_{i2}) + e_i. \quad (12)$$

Although these estimates are based on a misspecified model, if they are close to their true values then the partial plot is close to a plot of  $g(x_{i2}) + e_i$  versus  $x_{i2}$ .

Mansfield and Conerly (1987) considered the expectation properties of LS based partial residual plots by obtaining algebraic representations of the partial residuals using the true model distributional properties. Based on these representations, they showed, among other conclusions, that if the correct model was fitted then the expected partial residual plot should be a linear function of  $x_{2i}$ . They also showed that when  $\mathbf{x}_2$  and  $g$  are both orthogonal to  $\mathbf{X}_1$ , then we expect the partial residuals to be the unknown function  $g(x)$ . On the other hand, if  $\mathbf{x}_2$  and  $\mathbf{X}_1$  are highly collinear then there is little information in the partial residual plot.

Using the first-order approximation theory for robust residuals and fitted values established in McKean, Sheather and Hettmansperger (1990, 1993), McKean and Sheather (1997) obtained representations for the partial residuals when the true model is (9). Based on these representations, the conclusions described above of Mansfield and Conerly (1987) hold for the robust partial residual plots, also.

McKean and Sheather (1997) further developed a measure of efficiency between the robust and LS partial residual plots. If the correct model is fit then as discussed above the partial residual plot is expected to be the linear function  $\beta_2 \mathbf{x}_2$ . Thus the plot of interest would be that of  $\hat{\beta}_{W2} x_{i2}$  versus  $x_{i2}$  overlaid on the partial residual plot. Hence, it is the precision in the linear predicted equation  $\hat{\beta}_{W2} x_{i2}$  of  $\beta_2 x_{i2}$  which is of interest in terms of efficiency. This relative efficiency measure is given by the usual asymptotic relative efficiency between a robust estimate and the corresponding LS estimate. For example, if the Wilcoxon based residual plots are used then this asymptotic relative efficiency is given by

$$e_{W,LS} = \frac{\tau^2}{\sigma^2}, \quad (13)$$

where  $\sigma^2$  is the variance of the errors and  $\tau$  is defined in expression (6). If the error distribution is normal then  $e_{W,LS} = .955$ . However, if the error distribution is heavier tailed than the normal distribution then this ratio can be quite large; see Hettmansperger (1991).

In comparing the LS and robust representations of the partial residuals, McKean and Sheather (1997) showed that the random part of the LS partial residuals has unbounded influence while the corresponding part for the robust partial residuals has bounded influence. One bad outlier, say  $\epsilon_i$ , not only distorts the  $i$ th residual but other cases, also, because the representation of the LS partial residual includes the unbounded term  $\mathbf{H}\epsilon$ , where  $\mathbf{H}$  is the projection matrix onto the column space of  $[\mathbf{X}_1 : \mathbf{x}_2]$ . On the other hand, this is not true of the robust partial residuals because in the respective, representational expansion of the robust partial residuals all terms are bounded. The examples and sensitivity study found in Sections 4 and 5 provide illustrations of the distortion of LS partial residual plots due to outliers.

### 3.2 Augmented partial residual plots

The misspecified part of model, (10), is a first-order approximation to  $g(x)$ . We can also crawl up the Taylor series expansion of  $g(x)$  to fit higher order polynomials. This was proposed by Mallows (1986) for second-order representations. In this case, we fit the second-order model,

$$y_i = \alpha + \beta'_1 \mathbf{x}_{1i} + \beta_2 x_{2i} + \beta_3 x_{2i}^2 + e_i. \quad (14)$$

Now, the partial residuals are  $\hat{e}_i^* = \hat{e}_i + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{2i}^2$ , where  $\hat{e}_i$  are the residuals from the fit of model (14). Mallows called the resulting plot of  $\hat{e}_i^*$  versus  $x_{2i}$  the **augmented partial residual plots**. Another plot of interest here is  $\hat{e}_i^*$  versus  $\hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{2i}^2$ , because if the quadratic model is correct this later plot will appear linear; see the discussion above on the expected behavior of partial residual plots when the correct model is fit. Augmented plots are shown in Examples 1 and 3. Certainly higher degree polynomial approximations to  $g(x)$  can be handled in the same way as these quadratic plots.

## 4 Examples

In this section, we discuss several examples. We have chosen them to illustrate the exploratory behavior of the partial residual plots based on robust estimates and to show the sensitivity of the LS based partial residual plots to outlying observations. The data for all the examples is simulated, so at all times the correct model is known. There was little difference between the Wilcoxon based and the Huber based partial residual plots, so in a few examples only the results for the Wilcoxon based plots are shown. The Gauss-Newton type algorithm of Kapenga et al. (1988) was used to compute the Huber and Wilcoxon estimates.

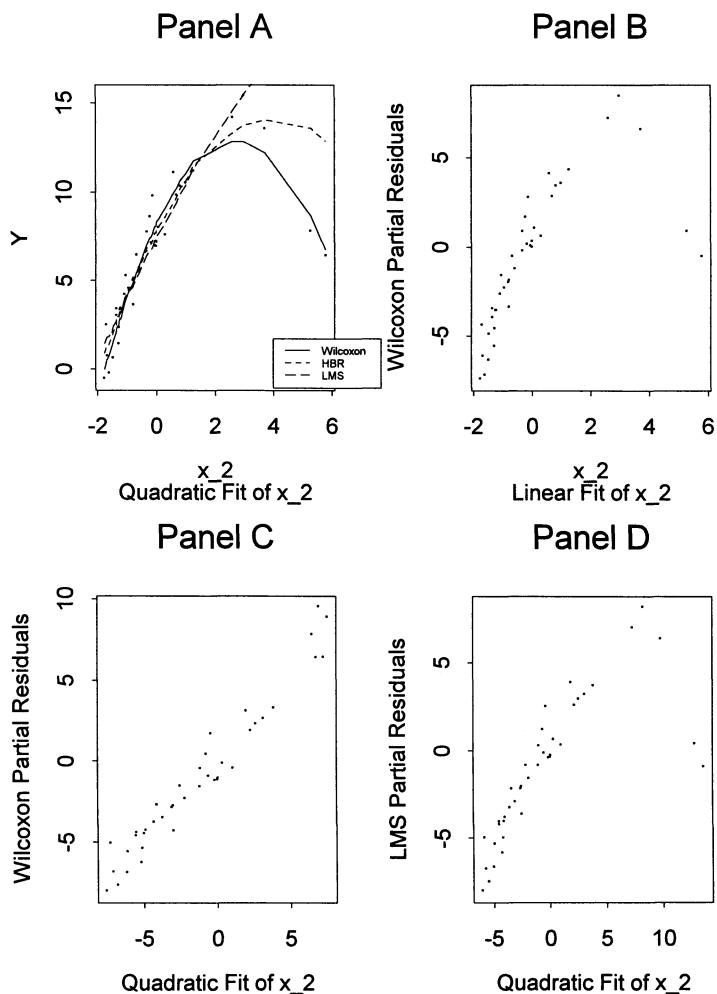


Figure 1: Plots for Example 1: Pane A, Data Overlaid with Wilcoxon, LMS and HBR fits; Panel B, Partial Residual Plot of the Wilcoxon Fit; Panel C, Augmented Partial Residual Plot of the Wilcoxon Fit; Panel D, Augmented Partiel Residual Plot of the LMS Fit.

### Example 1 *Quadratic Model*

The purposes of the first example is to show how the partial and augmented partial residual plots based on a robust fit behave for a simple quadratic model. It also shows why caution is necessary when considering partial residual plots based on high breakdown estimates. The generated

data follow the model

$$Y_i = 0 \cdot x_{1i} + 5.5|x_{i2}| - .6x_{i2}^2 + \epsilon_i , \quad (15)$$

where  $x_{i1}$  are iid uniform( $-1, 1$ ) variates, the  $\epsilon_i$ 's are simulated iid  $N(0, 1)$  variates and the  $x_{i2}$ 's are simulated contaminated normal variates with the contamination proportion set at .25 and the ratio of the variance of the contaminated part to the non-contaminated part set at 16. This was similar to an example discussed in Chang et al. (1997). Panel A of Figure 1 displays the scatterplot of the data overlaid by the Wilcoxon fit and two 50% breakdown fits: least median squares, LMS (Rousseeuw and Leroy, 1987), and a 50% high breakdown R estimate proposed by Chang et al. (1997), HBR. The LMS was computed using Stromberg's (1993) algorithm. Note that the fit based on the Wilcoxon estimates fits the curvature in the data quite well while the 50% breakdown estimates miss the curvature.

For the misspecified model

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 |x_{i2}| + e_i ,$$

Panel B of Figure 1 displays the partial residual plot based on the Wilcoxon fit. The plot clearly shows the need to fit a quadratic model. Panel C shows the augmented Wilcoxon partial residual plot when a quadratic model was fit. This is a plot of the partial residual versus the fit of the quadratic part. If the correct model has been specified then, as noted above, this plot should show a linear pattern, which it does. Panel D shows the same plot as Panel C except the fit based on the LMS estimates was used. Instead of a linear pattern, it shows a quadratic pattern, which is not helpful here because a quadratic model was fit.

### **Example 2** *Cook's (1993) Nonlinear Model*

This is an example discussed in Cook (1993). The observations are generated by

$$y_i = x_{i1} + x_{i2} + \frac{1}{1 + \exp(-x_{i3})} , \quad i = 1, \dots, 100 , \quad (16)$$

where  $x_{3i}$ 's are iid uniform(1, 26) random variables,  $x_{1i} = x_{3i}^{-1} + z_{1i}$ ,  $x_{2i} = \log x_{3i}^{-1} + z_{2i}$ ,  $z_{1i}$  has a  $N(0, .1^2)$  distribution,  $z_{2i}$  has a  $N(0, .25^2)$  distribution, and the  $z_{1i}$ 's and the  $z_{2i}$ 's are independent. A plot of the function  $g(x_{3i}) = \frac{1}{1 + \exp(-x_{i3})}$  versus  $x_{3i}$  appears in Panel A of Figure 2.

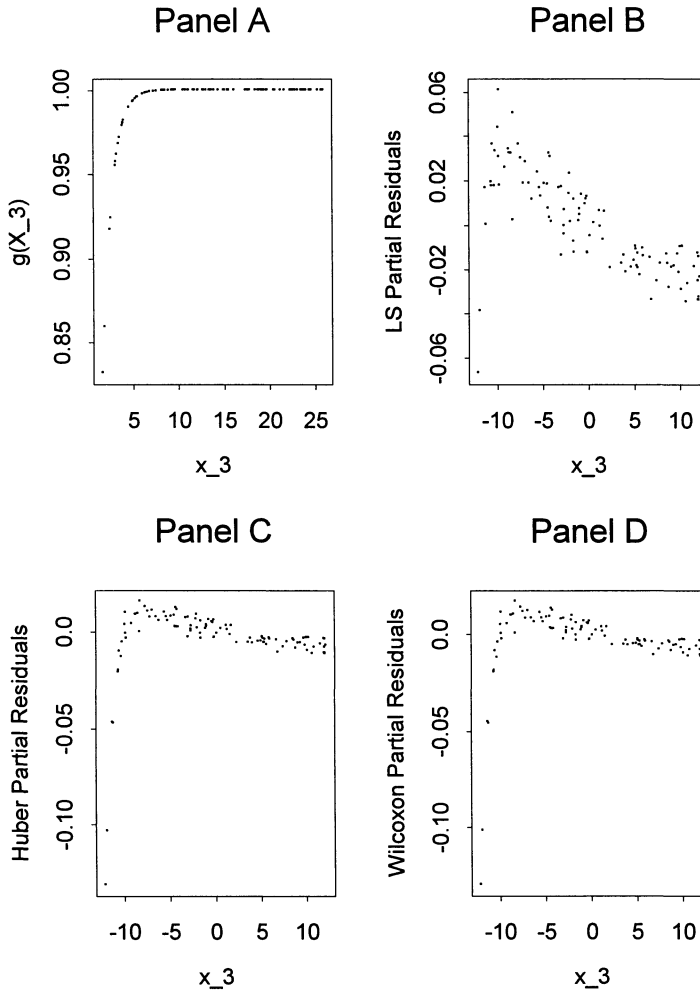


Figure 2: Plots for Example 2: Panel A, Plot of  $g(x_3)$  versus  $x_3$ ; Panel B, Partial Residual Plot of the LS Fit; Panel C, Partial Residual Plot of the Huber Fit; Panel D, Partial Residual Plot of the Wilcoxon Fit.

This is the function that the partial residual plots are attempting to identify. Panels B, C, and D display the partial residual plots based on the LS-, Huber and Wilcoxon fits, respectively. Note that the variable  $x_{i3}$  has been centered in these plots. The function  $g$  is identifiable from both robust residual plots, but  $g$  is not identifiable from the LS-plot. The points which distorted the LS partial residual plot are the points corresponding to the low values of  $x_{i3}$ . These points acted as outliers in  $Y$ -space and corrupted the LS fit, resulting in the poor LS partial residual plot. On the other

hand, the robust fits are much less sensitive to outliers in the  $Y$ -space than the LS fit; hence, these points did not corrupt the robust partial residual plots. As discussed in Section 3, one way of measuring efficiency in these plots is by the estimates of the constants of proportionality for the fitting procedure. For this example, these estimates are:  $\hat{\sigma} = .0162$ ,  $\hat{\kappa} = .0048$ , and  $\hat{\tau} = .0053$ . Hence, the robust estimates are three times more precise than the LS estimates on this data set.

Cook (1993) expands partial residual plots to the larger class of CERES plots. This procedure uses a nonparametric estimate of  $E(\mathbf{x}_1 | x_2)$  in place of the linear function  $\beta_2 x_{i2}$  in the regular partial residual plot in its construction of a partial residual plot. For this data set, as shown in the article by Cook, the procedure worked well with the LS fit. Similar plots could be developed based on robust fits, but for this example they are not needed.

### Example 3 *Berk and Booth's Model*

This is an example discussed in Berk and Booth (1995). The first-order partial residual plots fail on this example for all three fits. We include it, to show the importance of the augmented residual plot.

The values for the  $x_{i2}$ 's are the 100 values:  $-.99, -.97, \dots, .99$ . Then  $x_{i1}$  is generated as  $x_{i1} = x_{i2}^2 + .05z_{i1}$ , where  $z_{i1}$  are iid standard normal variates. The responses are generated by

$$y_i = x_{i2}^2 + .1z_{i2} , \quad (17)$$

where  $z_{i2}$ 's are iid standard normal variates and are independent of the  $z_{i1}$ 's. In this example,  $g(x_{i2}) = x_{i2}^2$  and Panel A of Figure 3 shows a plot of it versus  $x_{i2}$ . This is the function that the partial residual plots are attempting to identify. Panels B, C, and D display the partial residual plots based on the LS-, Huber and Wilcoxon fits, respectively. Note that none of them identify the function  $g$ . This is hardly surprising. The generating equation for  $x_{i1}$  is a strong quadratic in  $x_{i2}$ , there is little noise. Fitting  $x_{i1}$  stole the "clout" of  $x_{i2}$ . Also, the quadratic function  $g$  is centered over the region of interest. In its Taylor series expansion about 0, the linear term would not be important; hence, the inclusion of  $x_{i2}$  as linear will not help. If we crawl up the Taylor series expansion to include a second-order term then both of these conditions are alleviated and the (augmented) partial residual plot should identify the quadratic function. This is the case as demonstrated by Panels E and F of Figure 3, which are the augmented partial residual plots of the LS and Wilcoxon partial residuals versus the quadratic fit. The linearity of these plots indicate that the appropriate model has been fit.



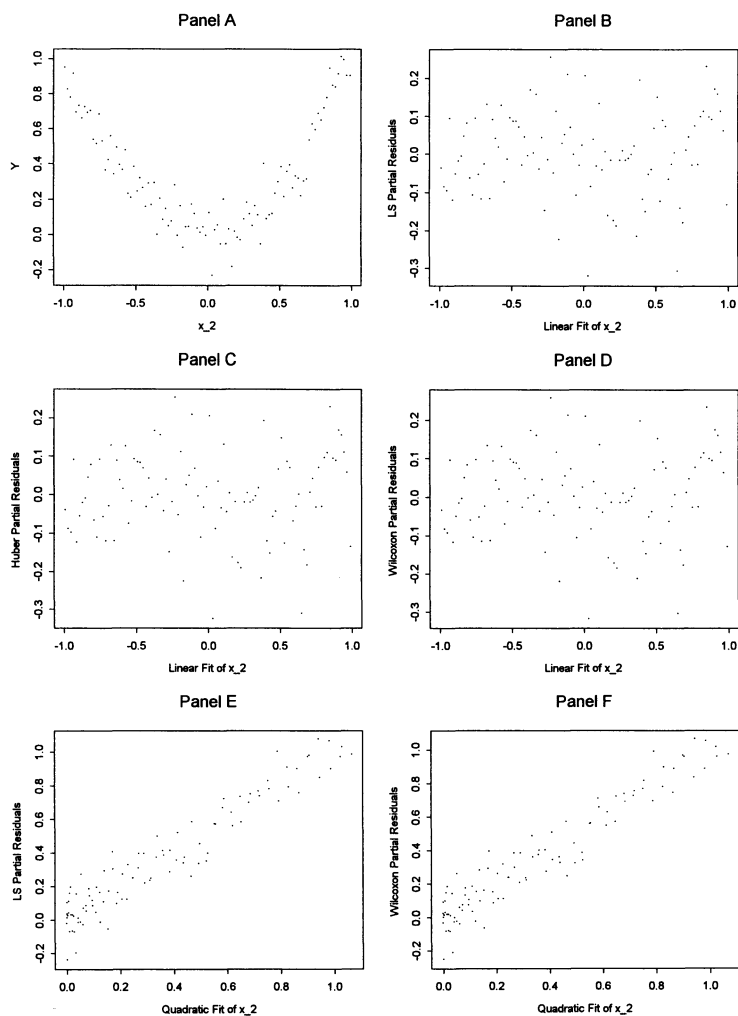


Figure 3: Plots for Example 3: Panel A, Plot of  $g(x_3)$  versus  $x_3$ ; Panel B, Partial Residual Plot of the LS Fit; Panel C, Partial Residual Plot of the Huber Fit; Panel D, Partial Residual Plot of the Wilcoxon Fit; Panel E, Augmented Partial Residual Plot of the LS Fit; Panel F, Augmented Partial Residual Plot of the Wilcoxon Fit.

## 5 Sensitivity Study

The following sensitivity study serves to illustrate the distortion of LS based partial residual caused by outliers in the  $Y$ -space. As our baseline model we consider a cubic polynomial in  $x_{i2}$  with normal errors. We chose  $x_{i1}$  to

be  $\text{uniform}(-1, 1)$  variates. The model is

$$y_i = 0 \cdot x_{i1} + 5.5x_{i2}^2 - .6x_{i2}^3 + \epsilon_i ; i = 1, \dots, 40 , \quad (18)$$

where  $\epsilon_i$  are iid  $N(0, 1)$  and  $x_{i2}$  are generated from a contaminated normal distribution. The misspecified model is

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i . \quad (19)$$

In this setting, the partial residual plots should easily show that a cubic needs to be fit. This is confirmed by the top row of Figure 4 which are the partial residual plots based on the LS and the Wilcoxon fits, respectively, when the misspecified model, (19), is fitted.

Next, in a series of four stages we distorted the values of three of the responses, as shown in Table 1, from small to large changes of these values. We then obtained the partial residual plots based on the LS and Wilcoxon fits for each of these stages.

Table 1: Successive changes to the response variable for Model.

| Case | Original | Stage 1 | Stage 2 | Stage 3 | Stage 4 |
|------|----------|---------|---------|---------|---------|
| 3    | -.567    | 5.567   | 10.56   | 100.56  | 1000.56 |
| 33   | 62.44    | 310.44  | -620.44 | -620.44 | -6200.4 |
| 40   | 67.24    | -335.2  | -670.2  | 670.2   | -6700.2 |

Column A of Figure 4 shows the effect these changes had on the LS partial residual plot. Note that limit on the vertical axes were changed so that the bulk of the cases could be plotted. The distortion is obvious. From a clear cubic pattern for the baseline model (the top row of the plots) the LS based partial residual plot becomes more and more distorted as the successive stages are fitted. The clear cubic pattern has been lost even in the first stage (the second row of the plots). By the second stage (third row) the cubic pattern is not identifiable. There is some linear trend in the third stage (fourth row), but in the final stage (last row) there is just noise. On the other hand, the cubic pattern is clearly identifiable in the robust partial residual plots (Column B of Figure 4) for all stages.

## 6 Conclusion

LS partial residual plots are an important diagnostic tool for exploratory fitting. They are often used to identify unknown functions of the predictors. They are, however, vulnerable to the effect of outliers. One large outlier can severely distort the LS based partial residual plot, making the identification of the unknown function of the predictor difficult to impossible.

Furthermore because the function  $g$  can be nonlinear, good data can have the same effect on the LS partial residual plots as outliers; see Example 2.

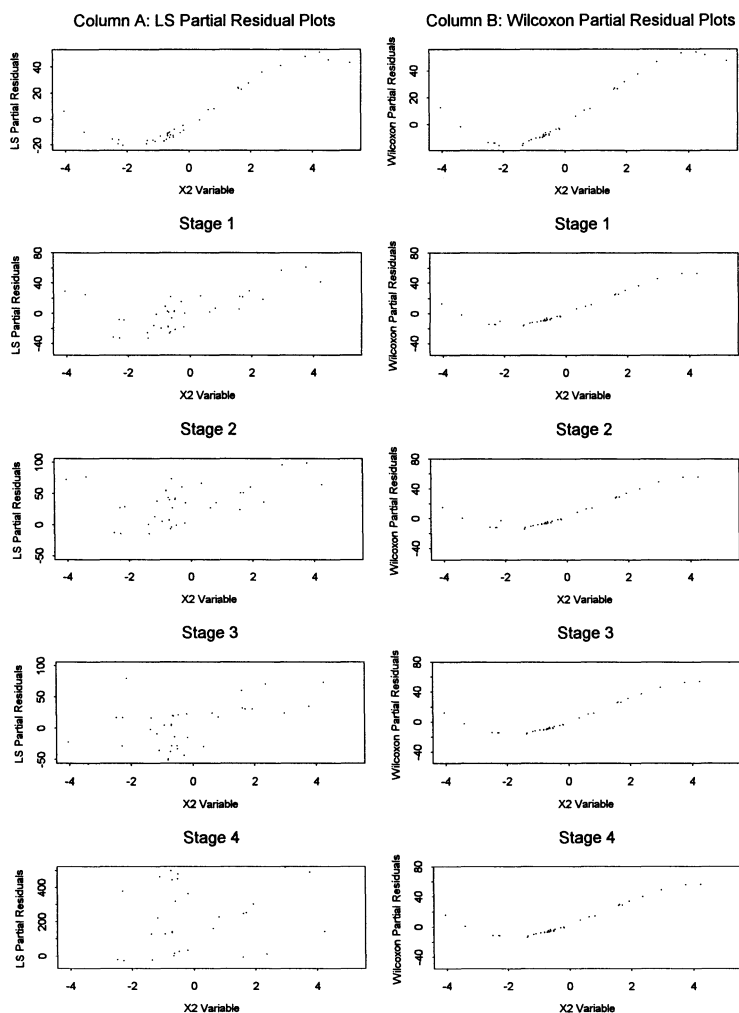


Figure 4: Plots for Sensitivity Study : Column A, LS Partial Residual Plots for Original Data Followed by the LS Plots for Stages 1-4; Column B, Wilcoxon Partial Residual Plots for Original Data Followed by the Wilcoxon Plots for Stages 1-4.

In this paper, we have presented partial residual plots based on robust estimates. As the examples and sensitivity study demonstrated these partial residual plots are effective in exploratory fitting. Furthermore they are not vulnerable to the effect of outliers as their LS counterparts. Also for highly nonlinear situations such as Example 2, they are able to easily

identify the unknown function. As the sensitivity study shows, even in the presence of severe outlying observations partial residual plots based on highly efficient robust estimates are able to retain their focus, making the identification of the unknown functions possible.

## References

- [1] Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The NEW S Language*. New York: Chapman and Hall.
- [2] Berk, K. N. and Booth, D. E. (1995). Seeing a curve in multiple regression. *Technometrics* **37** 385-398.
- [3] Chang, W., McKean, J. W., Naranjo, J. D. and Sheather, S. J. (1997). High breakdown rank regression. Submitted.
- [4] Cook, R. D. (1993). Exploring partial residual plots. *Technometrics* **35** 351-362.
- [5] Cook, R. D., Hawkins, D. M. and Weisberg, S. (1992). Comparison of model misspecification diagnostics using residuals from least mean of squares and least median of squares fits. *J. Am. Statist. Assoc.* **87** 419-424.
- [6] Ezekiel, M. (1924). A Method for Handling Curvilinear Correlation for any Number of Variables. *J. Am. Statist. Assoc.* 431-453.
- [7] Hettmansperger, T. P. (1991). *Statistical Inference Based on Ranks*. New York: Kruger.
- [8] Hettmansperger, T. P., McKean, J. W. and Sheather, S. J. (1997). Rank-based analyses of linear models. In *Handbook of Statistics*, Ed. C. R. Rao, Vol. 15. In press.
- [9] Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.
- [10] Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.* **43** 1449-1458.
- [11] Kapenga, J. A., McKean, J. W. and Vidmar, T. J. (1988). *RGLM: Users Manual*. Amer. Statist. Assoc. Short course on robust statistical procedures for the analysis of linear and nonlinear models. New Orleans.
- [12] Koul, H. L. Sievers, G. L. and McKean, J. W. (1987). An estimator of the scale parameter for the rank analysis of linear models under general score functions. *Scandi. J. Statist.* **14** 131-141.
- [13] Larsen, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics* **14** 781-790.
- [14] Mallows, C. L. (1986). Augmented partial residual plots. *Technometrics* **28** 313-320.
- [15] Mansfield, R. E. and Conerly, M. D. (1987). Diagnostics value of residual and partial residual plots. *Am. Statistician* **41** 107-116.

- [16] McKean, J. W. Naranjo, J. D. and Sheather, S. J. (1996a). Diagnostics to detect differences in robust fits of linear models. *Comput. Statist.* **11** 223-243.
- [17] McKean, J. W. Naranjo, J. D. and Sheather, S. J. (1996b). An efficient and high breakdown procedure for model criticism. *Commun. Statist. Theory and Methods* **25** 2575-2595.
- [18] McKean, J. W. and Schrader, R. M. (1984). A comparison of methods for studentizing the sample median. *Commun. Statist. B* **6** 751-773.
- [19] McKean, J. W. Sheather, S. J. (1997). Partial residual plots based on robust fitting. Submitted.
- [20] McKean, J. W. Sheather, S. J. and Hettmansperger, T. P. (1990). Regression diagnostics for rank-based methods. *J. Am. Statist. Assoc.* **85** 1018-28.
- [21] McKean, J. W. Sheather, S. J. and Hettmansperger, T. P. (1993). The use and interpretation of residuals based on robust estimation. *J. Am. Statist. Assoc.* **88** 1254-1263.
- [22] McKean, J. W. Sheather, S. J. and Hettmansperger, T. P. (1994). Robust and high breakdown fits of polynomial models. *Technometrics* **36** 409-415.
- [23] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
- [24] Stromberg, A. J. (1993). Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM J. Scient. Comput.* **14** 1289-1299.

# Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness

Elias Moreno

*University of Granada, Spain*

*Abstract:* For model selection the Bayes factor is not well defined when using default priors since they are typically improper. To overcome this problem two methods have recently been proposed. These methods, intrinsic and fractional, are studied here as methods to producing proper prior distributions for model selection from the improper conventional priors for estimation. For nested models, fractional priors are here defined and a comparison with intrinsic priors introduced by Berger and Pericchi is carried out. Robustness of the Bayes factor as the prior varies over the classes of intrinsic and fractional priors, is studied. Some illustrative examples are provided.

*Key words:* Bayes factor, bayesian robustness, fractional priors, intrinsic priors, model selection.

AMS subject classification: Primary 62A15; secondary 62F15.

## 1 Introduction

Suppose that two models  $M_1$  and  $M_2$  are proposed to describe the data  $z = (x_1, x_2, \dots, x_n)$ . Under model  $M_i$  the data are distributed as  $f_i(z|\theta_i)$ , and the prior distribution for  $\theta_i$  is  $\pi_i(\theta_i)$ ,  $i = 1, 2$ . The Bayesian way to compare the two models consists in computing the posterior odds

$$\frac{\Pr(M_2|z)}{\Pr(M_1|z)} = B_{21}(z) \frac{\Pr(M_2)}{\Pr(M_1)}.$$

Thus, the Bayes factor  $B_{21}(z)$  encapsulates all what the data have to say

about such a comparison. This Bayes factor is given by

$$B_{21}(z) = \frac{\int_{\Theta_2} f_2(z|\theta_2)\pi_2(\theta_2)d\theta_2}{\int_{\Theta_1} f_1(z|\theta_1)\pi_1(\theta_1)d\theta_1}.$$

To subjectively elicit the priors  $\pi_i(\theta_i)$ ,  $i = 1, 2$ , is most cases a very difficult task. A way to alleviate this task is to elicit instead of a single prior for each model, a class of prior distributions  $\Gamma = \{\pi_1(\theta_1), \pi_2(\theta_2)\}$  that maintains the features of the priors on which we are confident. Hence, the Bayes factors as the prior ranges over  $\Gamma$  takes now values in the range

$$(\inf_{\Gamma} B_{21}(z), \sup_{\Gamma} B_{21}(z)).$$

This range is generally too large and it typically gives  $\inf_{\Gamma} B_{21}(z) = 0$ , so that there are priors in  $\Gamma$  favouring  $M_1$  and also prior favouring  $M_2$ . Hence, it does not allow to decide which of the model is supported by the data.

Another way to deal with the problem of model selection is to set as  $\pi_i(\theta_i)$  the conventional prior for estimation of  $\theta_i$ , say  $\pi_i^N(\theta_i)$ , which typically is improper, that is the integral  $\int_{\Theta_i} \pi_i^N(\theta_i)d\theta_i$  diverges. This means that no normalization of  $\pi_i^N(\theta_i)$  is possible so that it is defined up to an arbitrary multiplicative constant. This implies that  $B_{21}(z)$  is defined up to a ratio of unspecified constants.

There are in the literature several ways either to specify the constants or to remove them from the analysis, see Akaike (1973), Schwarz (1978), Spiegelhalter and Smith (1982), O'Hagan (1995), Berger and Pericchi (1995, 1996), among others. For a recent review, see Kass and Raftery (1995).

In this paper we focus on intrinsic and fractional methodologies as methods to producing proper prior distributions for model comparison. This is motivated by the fact that while there are methods to produce prior for estimation that work reasonably well, there is a lack of such a methods for model selection.

Let us briefly summarize the intrinsic and fractional Bayes factors. The *intrinsic Bayes factor* (IBF) was proposed by Berger and Pericchi (1995, 1996). This is a partial Bayes factor based on a minimum training sample, say  $x(l)$ , which is a minimal subsample of the sample  $z$  such that  $0 < \int_{\Theta_i} f_i(x(l)|\theta_i)\pi_i^N(\theta_i)d\theta_i < \infty$ ,  $i = 1, 2$ . This part of the sample is devoted to convert  $\pi_i^N(\theta_i)$  into  $\pi_i^N(\theta_i|x(l))$ , which is now proper, and the rest of the data is devoted to construct the intrinsic Bayes factor for model comparison, using the  $\pi_i^N(\theta_i|x(l))$  as priors. Thus the partial Bayes factor is defined as

$$B_{21}^I(x(-(l)|x(l))) = B_{21}^N(z) B_{12}^N(x(l))$$

where  $B_{21}^N(z)$  is the Bayes factor for the improper priors  $\pi_i^N(\theta_i)$ ,  $i = 1, 2$ , and sample  $z$ . Note that  $B_{21}^I(x(-(l)|x(l)))$  does not depend on the arbitrary constant involved in the improper priors  $\pi_i^N(\theta_i)$ ,  $i = 1, 2$ . To avoid the dependence of the partial Bayes factor on the particular  $x(l)$ , they introduced an average on the set of all training samples. Thus, the *arithmetic intrinsic* Bayes factor is defined as

$$B_{21}^{AI}(z) = B_{21}^N(z) \frac{1}{L} \sum_{l=1}^L B_{12}^N(x(l)), \quad (1)$$

where  $L$  is the number of training sample contained in  $z$ .

The fractional method, proposed by O'Hagan, considers the *fractional Bayes factor* (FBF) which, as the author motivates, *is defined by analogy with the partial Bayes factor to avoid the arbitrariness of choosing a particular training sample*. The FBF is defined as

$$B_{21}^b(z) = B_{21}^N(z) \frac{\int_{\Theta_1} f_1(z|\theta_1)^b \pi_1^N(\theta_1) d\theta_1}{\int_{\Theta_2} f_2(z|\theta_2)^b \pi_2^N(\theta_2) d\theta_2} \quad (2)$$

where  $b$  is a constant that depends on the sample size  $n$ . Notice that  $B_{21}^b(z)$  does not depend on the arbitrary constants involved in the improper priors.

Both, the IBF and the FBF, contain  $B_{21}^N(z)$  as a common factor. The other factors appearing in the right hand side of (1) and (2) can be considered as the correction term of  $B_{21}^N(z)$  to avoid the dependence of the unspecified constants. These correction terms are different. Furthermore, while the IBF correction term is completely specified, the FBF correction term depends on  $b$  that has to be assessed. We will go back on this topic in Subsection 2.4.

A crucial property of the intrinsic method is that it is capable to generating proper priors. These priors are derived by imposing that the IBF is asymptotically equivalent to an actual Bayes factor for the so-called *intrinsic priors* (see Berger and Pericchi, 1996, for the definition of *intrinsic priors* and Moreno, Bertolino and Racugno, 1996, for a characterization). In some sense this guarantee that the IBF is a truly Bayes factor.

The question is if something similar can be said on the FBF. In Subsection 2.2 we produce a functional equation to derive *fractional priors* that enjoy the same spirit of the *intrinsic priors*. The solution of this equation is also discussed. Computation of the Bayes factor for intrinsic priors entails a robustness issue since they are not unique, see Subsection 2.1. It will be also shown that the solution to the fractional equation is not unique so that a similar robustness issue appears here. Robustness is studied in Subsection 2.3. In Section 3 some illustrative examples are given. Section 4 contains some conclusions.



## 2 The intrinsic and fractional priors for nested models

Suppose that the two sampling models under comparison  $\{f_1(x|\theta_1), \theta_1 \in \Theta_1\}$ ,  $\{f_2(x|\theta_2), \theta_2 \in \Theta_2\}$  are nested. This means that the following conditions are satisfied,

- (i)  $\Theta_1 \subset \Theta_2$ ,
- (ii)  $f_2(x|\theta_2) \equiv f_1(x|\theta_1)$ , for  $\theta_2 = \theta_1$ .

Let  $\pi_i^N(\theta_i)$ ,  $i = 1, 2$ , the improper priors chosen.

### 2.1 The intrinsic priors

The *intrinsic priors* are shown to be, see Moreno, Bertolino and Racugno (1996). any pair  $(\pi_1(\theta_1), \pi_2(\theta_2))$  such that,

- (a)  $\pi_1(\theta_1)$  is any prior in the class

$$\Gamma_1 = \{\pi_1(\theta_1) : \int_{\Theta_1} \pi_1(\theta_1) d\theta_1 = 1, \int_{\Theta_2} T(\theta_2) \pi_1(\psi_1(\theta_2)) d\theta_2 = 1\}$$

where  $T(\theta_2) = \frac{\pi_2^N(\theta_2)}{\pi_1^N(\psi_1(\theta_2))} E_{x(l)|\theta_2}^{M_2} B_{12}^N(x(l))$  and  $\psi_1(\theta_2)$  is the limit point of the MLE  $\hat{\theta}_1(z)$  for parameter  $\theta_1$  in  $M_1$  when sampling from  $M_2$  at point  $\theta_2$ .

- (b) For each  $\pi_1(\theta_1) \in \Gamma_1$ ,  $\pi_2(\theta_2)$  is given by

$$\pi_2(\theta_2) = T(\theta_2) \pi_1(\psi_1(\theta_2)).$$

For studying robustness of the Bayes factor with respect to intrinsic priors it is convenient to express  $\Gamma_1$  as

$$\Gamma_1 = \{\pi_1(\theta_1) : \int_{\Theta_1} \pi_1(\theta_1) d\theta_1 = 1, \int_{\Theta_1} V(\theta_1) \pi_1(\theta_1) d\theta_1 = 1\},$$

where  $V(\theta_1) = T(\theta_1) + \{\int_{\Theta_2(\theta_1)} T(\theta_2) d\theta_2\} 1_{\Theta_1^*}(\theta_1)$ ,  $\Theta_2(\theta_1) = \{\theta_2 \in (\Theta_2 - \Theta_1) : \psi_1(\theta_2) = \theta_1\}$ , the  $\psi_1$ -coset of  $\theta_1$  in  $\Theta_2 - \Theta_1$ , and  $\Theta_1^*$  is the  $\psi_1$ -image of  $\Theta_2 - \Theta_1$ .

### 2.2 The fractional priors

Let us first precisely state what we mean by *fractional priors*.

**Definition 1** For the sampling models  $\{f_i(x|\theta_i), i = 1, 2\}$  the proper priors  $\pi_1(\theta_1), \pi_2(\theta_2)$  are called *fractional priors* if its Bayes factor and the FBF for  $\pi_i^N(\theta_i)$ ,  $i = 1, 2$ , are asymptotically equivalent for some sequence  $\{b_n\}$ .

The following condition is necessary in order to propose the fractional equation from which the fractional priors are derived.

**Assumption (A)** The nested models  $\{f_i(x|\theta_i), \pi_i^N(\theta_i), i = 1, 2\}$  satisfy Assumption (A) if for some sequence  $\{b_n\}$  the limit in probability  $[P_{\theta_2}]$  of the correction term of the FBF is a degenerated random variable. In other words, under Assumption (A) there exists a function  $F_{12}^{M_2}(\theta_2)$ , which could be a constant, such that

$$F_{12}^{M_2}(\theta_2) = \lim_{n \rightarrow \infty} [P_{\theta_2}] \frac{\int_{\Theta_1} f_1(z|\theta_1)^{b_n} \pi_1^N(\theta_1) d\theta_1}{\int_{\Theta_2} f_2(z|\theta_2)^{b_n} \pi_2^N(\theta_2) d\theta_2}$$

For simplicity in notation the dependence of  $F_{12}^{M_2}(\theta_2)$  on the sequence  $\{b_n\}$  is not made explicitly.

**Theorem 1** *Under Assumption (A), the fractional priors  $(\pi_1(\theta_1), \pi_2(\theta_2))$  are the solutions to the functional equation*

$$F_{12}^{M_2}(\theta_2) = \frac{\pi_2(\theta_2)}{\pi_2^N(\theta_2)} \frac{\pi_1^N(\psi_1(\theta_2))}{\pi_1(\psi_1(\theta_2))}. \quad (3)$$

**Proof:** If we expand  $\frac{\pi_i(\theta_i)}{\pi_i^N(\theta_i)}$  around the MLE  $\hat{\theta}_i(z)$ , the Bayes factor for the fractional priors  $\pi_1(\theta_1), \pi_2(\theta_2)$ , can be approximated as

$$\begin{aligned} B_{21}(z) &= \frac{\int_{\Theta_2} f_2(z|\theta_2) \pi_2^N(\theta_2) \frac{\pi_2(\theta_2)}{\pi_2^N(\theta_2)} d\theta_2}{\int_{\Theta_1} f_1(z|\theta_1) \pi_1^N(\theta_1) \frac{\pi_1(\theta_1)}{\pi_1^N(\theta_1)} d\theta_1} \\ &= B_{21}^N(z) \frac{\pi_2(\hat{\theta}_2(z))}{\pi_2^N(\hat{\theta}_2(z))} \frac{\pi_1^N(\hat{\theta}_1(z))}{\pi_1(\hat{\theta}_1(z))} (1 + o(1)). \end{aligned}$$

Equating the limit in probability  $[P_{\theta_2}]$  of the fractional Bayes factor given in (2) with the limit in probability  $[P_{\theta_2}]$  of the above expression, we obtain

$$F_{12}^{M_2}(\theta_2) = \lim_{n \rightarrow \infty} [P_{\theta_2}] \frac{\pi_2(\hat{\theta}_2(z))}{\pi_2^N(\hat{\theta}_2(z))} \frac{\pi_1^N(\hat{\theta}_1(z))}{\pi_1(\hat{\theta}_1(z))},$$

where the left hand side follows from Assumption (A). Notice that we do not need to take limit in probability under model  $M_1$  since it is nested in  $M_2$ . This gives (3) and proves the assertion.  $\square$

We remark that the fractional priors does not depends on the arbitrary constants involved in the improper priors. They cancel out in expression (3).

**Corollary 1** *Fractional priors are any pair  $(\pi_1(\theta_1), \pi_2(\theta_2))$ , where  $\pi_1(\theta_1)$  is any member of the class*

$$\Gamma_2 = \{\pi_1(\theta_1) : \int_{\Theta_1} \pi_1(\theta_1) d\theta_1 = 1, \int_{\Theta_2} S(\theta_2) \pi_1(\psi_1(\theta_2)) d\theta_2 = 1\},$$

with

$$S(\theta_2) = F_{12}^{M_2}(\theta_2) \frac{\pi_2^N(\theta_2)}{\pi_1^N(\psi_1(\theta_2))},$$

and for each  $\pi_1(\theta_1) \in \Gamma_2$ ,

$$\pi_2(\theta_2) = S(\theta_2) \pi_1(\psi_1(\theta_2)).$$

**Proof:** Equation (3) can be written as

$$\pi_2(\theta_2) = S(\theta_2) \pi_1(\psi_1(\theta_2)),$$

where  $\pi_1(\theta_1)$  have to be a probability distribution such that  $\pi_2(\theta_2)$  be also a probability distribution. This proves Corollary 1.  $\square$

It is convenient to rewrite class  $\Gamma_2$  as

$$\Gamma_2 = \{\pi_1(\theta_1) : \int_{\Theta_1} \pi_1(\theta_1) d\theta_1 = 1, \int_{\Theta_1} H(\theta_1) \pi_1(\theta_1) d\theta_1 = 1\},$$

where

$$H(\theta_1) = S(\theta_1) + \left\{ \int_{\Theta_2(\theta_1)} S(\theta_2) d\theta_2 \right\} 1_{\Theta^*(\theta_1)}(\theta_1).$$

### 2.3 Robustness of the Bayes factor for intrinsic and fractional priors

From Subsection 2.1 it follows that the Bayes factor for the *intrinsic priors*  $(\pi_1(\theta_1), \pi_2(\theta_2))$  is

$$B_{21}(z) = \frac{\int_{\Theta_2} f_2(z|\theta_2) T(\theta_2) \pi_1(\psi_1(\theta_2)) d\theta_2}{\int_{\Theta_1} f_1(z|\theta_1) \pi_1(\theta_1) d\theta_1},$$

which is written only in term of  $\pi_1(\theta_1)$ .

It is easily shown that  $B_{21}(z)$ , can be expressed as

$$B_{21}(z) = \frac{\int_{\Theta_1} W(z; \theta_1) \pi_1(\theta_1) d\theta_1}{\int_{\Theta_1} f_1(z|\theta_1) \pi_1(\theta_1) d\theta_1}$$

where  $\pi_1(\theta_1) \in \Gamma_1$  and

$$W(z; \theta_1) = f_2(z|\theta_1) T(\theta_1) + \left\{ \int_{\Theta_2(\theta_1)} f_2(z|\theta_2) T(\theta_2) d\theta_2 \right\} 1_{\Theta^*(\theta_1)}(\theta_1).$$

On the other hand, from Corollary 1 it follows that the Bayes factor for fractional priors  $\tilde{B}_{21}(z)$  can be written only in term of  $\pi_1(\theta_1) \in \Gamma_2$  as

$$\tilde{B}_{21}(z) = \frac{\int_{\Theta_2} f_2(z|\theta_2)S(\theta_2)\pi_1(\psi_1(\theta_2))d\theta_2}{\int_{\Theta_1} f_1(z|\theta_1)\pi_1(\theta_1)d\theta_1}.$$

It is straightforward to show that this Bayes factor associated to fractional priors can be written as

$$\tilde{B}_{21}(z) = \frac{\int_{\Theta_1} G(z; \theta_1)\pi_1(\theta_1)d\theta_1}{\int_{\Theta_1} f_1(z|\theta_1)\pi_1(\theta_1)d\theta_1},$$

where

$$G(z; \theta_1) = f_2(z|\theta_1)S(\theta_1) + \left\{ \int_{\Theta_2(\theta_1)} f_2(z|\theta_2)S(\theta_2)d\theta_2 \right\} 1_{\Theta^*(\theta_1)}(\theta_1).$$

Global robustness of the Bayes factors  $B_{21}(z)$  and  $\tilde{B}_{21}(z)$  as  $\pi_1$  ranges over  $\Gamma_1$  and  $\Gamma_2$  respectively, can be established by computing the ranges  $(\inf_{\pi_1 \in \Gamma_1} B_{21}(z), \sup_{\pi_1 \in \Gamma_1} B_{21}(z))$ ,  $(\inf_{\pi_1 \in \Gamma_2} \tilde{B}_{21}(z), \sup_{\pi_1 \in \Gamma_2} \tilde{B}_{21}(z))$ . This involve a moment problem for which Theorem 2 summarizes the solution.

**Theorem 2** *The infimum of the fractional Bayes factor as the priors range over the class of priors  $\Gamma_2$ , say  $\underline{\lambda} = \inf_{\pi_1 \in \Gamma_2} \tilde{B}_{21}(z)$ , is the unique solution in  $\underline{\lambda}$  to the equation*

$$\sup_{d \in R} \inf_{\theta_1 \in \Theta_1} [G(z; \theta_1) - \underline{\lambda} f_1(z|\theta_1) + d(1 - H(\theta_1))] = 0.$$

*The sup is obtained by interchanging in the above expression sup with inf. A similar statement can be given for the Bayes factor for the intrinsic priors.*

**Proof:** The proof follows by using the linearization algorithm, see for instance Lavine, Wasserman and Wolpert (1993), and the so-called Generalized Moment Theory, Kemperman (1987), Salinetti (1994) and Liseo, Moreno and Salinetti (1996).

For data  $z$ , robustness of a Bayes factor  $B_{21}(z)$  as the priors range over a given class  $\Gamma$ , is strictly achieved if either  $\sup_{\pi \in \Gamma} B_{21}(z) < 1$  favouring  $M_1$ , or  $\inf_{\pi \in \Gamma} B_{21}(z) > 1$ , in which case  $M_2$  is favoured. With obvious adaptations of the suggestion by Jeffreys (see Kass and Raftery, 1995) we would take the following interpretation:

- if  $1 < \inf_{\pi \in \Gamma_1} B_{21}(z) < \sqrt{10}$ , the evidence against  $M_1$  is small,
- if  $\sqrt{10} < \inf_{\pi \in \Gamma_1} B_{21}(z) < 10$ , the evidence against  $M_1$  is substantial,
- if  $10 < \inf_{\pi \in \Gamma_1} B_{21}(z) < 100$ , the evidence against  $M_1$  is strong and
- if  $100 < \inf_{\pi \in \Gamma_1} B_{21}(z)$ , the evidence against  $M_1$  is decisive.

## 2.4 The role of the sequence $\{b_n\}$ in the fractional priors

An important difference between the intrinsic priors and the fractional priors is that while the former are derived automatically from the specification of the models  $\{f_i(x|\theta_i), \pi_i^N(\theta_i), i = 1, 2\}$ , the latter needs in addition to assess the sequence  $\{b_n\}$ . In fact, for producing fractional priors we already have a restriction on this sequence since Assumption (A) has to be satisfied. Nevertheless, this does not guarantee the uniqueness of the sequence, so that an additional convention has to be imposed. Let us illustrate the assertion with the following simple example.

**Example 1** *Consider the nested models*

$$M_1 : f_1(x|\theta_1) = N(x|\theta_1, 1), \quad \pi_1(\theta_1) = 1_{\{0\}}(\theta_1),$$

$$M_2 : f_2(x|\theta_2) = N(x|\theta_2, 1), \quad \pi_2^N(\theta_2) \propto 1_R(\theta_2),$$

that is, we are testing that the mean of a normal distributions is 0 versus it is different from 0. Notice that the prior for the first model is proper and the prior for the second is the conventional uniform improper prior.

The intrinsic priors for this models can be shown to be the unique pair

$$\pi_1(\theta_1) = 1_{\{0\}}(\theta_1), \quad \pi_2(\theta_2) = N(\theta_2|0, 2).$$

On the other hand the fractional priors are derived as follows. For a given sequence  $\{b_n\}$ , we have

$$\begin{aligned} F_{12}^{M_2}(\theta_2) &= \lim_{n \rightarrow \infty} [P_{\theta_2}] \frac{\int_{\Theta_1} f_1(z|\theta_1)^{b_n} \pi_1^N(\theta_1) d\theta_1}{\int_{\Theta_2} f_2(z|\theta_2)^{b_n} \pi_2^N(\theta_2) d\theta_2} \\ &= \lim_{n \rightarrow \infty} [P_{\theta_2}] \frac{\sqrt{n b_n}}{\sqrt{2\pi}} \exp\{-b_n \frac{n\bar{x}}{2}\}, \end{aligned}$$

where  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ . The sequences  $\{b_n\}$  proposed by O'Hagan (1995) are

$$b_n = \frac{m_0}{n}, \quad m_0 = 1, 2, \dots, n-1,$$

$$b'_n = \frac{\sqrt{m_0}}{n},$$

$$b''_n = \frac{\log n}{n}.$$

The sequences  $\{b'_n\}, \{b''_n\}$  do not satisfy Assumption (A), so that they do not produce fractional priors. Therefore, we are left with sequences of the form  $\{b_n = \frac{m_0}{n}, m_0 = 1, 2, \dots, n-1\}$ , for which we obtain

$$F_{12}^{M_2}(\theta_2) = \frac{\sqrt{m_0}}{\sqrt{2\pi}} \exp\{-m_0 \frac{\theta_2^2}{2}\} = N(\theta_2|0, \frac{1}{m_0}).$$

Thus, the *fractional priors* are

$$\pi_1(\theta_1) = 1_{\{0\}}(\theta_1), \quad \pi_2(\theta_2) = N(\theta_2|0, \frac{1}{m_0}), \quad m_0 = 1, \dots, n-1.$$

Note that  $\pi_2(\theta_2)$  is not unique but depends on  $m_0$ . The convention we will take is to fix  $m_0$  as the minimal training sample size (Berger and Mortera, 1995). In this case  $m_0 = 1$ . For this assessment the fractional prior for  $M_2$  is the density  $N(\theta_2|0, 1)$ .

### 3 Examples

Let us illustrate the behaviour of the Bayes factor for intrinsic and fractional priors for two standard problems. The first is one sided testing on the mean of a normal distribution and the second is a two sided testing.

**Example 2 One sided testing.** Let  $X$  be a random variable  $N(x|\theta, 1)$  distributed. Suppose that we are interested in testing  $H_1 : \theta \geq 0$  versus  $H_2 : \theta < 0$ .

A formulation of this testing problem in a nested context would be to compare the two models

$$\begin{aligned} M_1 : f_1(x|\theta_1) &= N(x|\theta_1, 1), \quad \pi_1^N(\theta_1) \propto 1_{[0, \infty)}(\theta_1), \\ M_2 : f_2(x|\theta_2) &= N(x|\theta_2, 1), \quad \pi_2^N(\theta_2) \propto 1_{(-\infty, \infty)}(\theta_2), \end{aligned}$$

where the priors are the standard improper prior for location parameter. It is easy to see that  $\psi_1(\theta_2) = \theta_2 1_{[0, \infty)}(\theta_2)$ .

The class of intrinsic priors is

$$\Gamma_1 = \left\{ \pi_1 : \int_0^\infty \Phi\left(\frac{\theta_1}{\sqrt{2}}\right) \pi_1(\theta_1) d\theta_1 = 1 - \frac{k}{\sqrt{\pi}}, 0 \leq k \leq \frac{1}{2}\sqrt{\pi} \right\},$$

where  $k$  is the value of  $\pi_1(\theta_1)$  at the discontinuity point  $\theta_1 = 0$  (see Moreno, Bertolino and Racugno, 1996) and  $\Phi(\theta_1)$  is the standard cumulative distribution function at point  $\theta_1$ . Thus, for a given sample  $(\bar{x}, n)$  and  $\pi_1(\theta_1) \in \Gamma$ , the Bayes factor for  $M_2$  against  $M_1$  can be shown to be

$$B_{21}(\bar{x}, n) = \frac{kN(\bar{x}, n) + \int_0^\infty \exp\left\{-\frac{n(\bar{x}-\theta_1)^2}{2}\right\} \Phi\left(\frac{\theta_1}{\sqrt{2}}\right) \pi_1(\theta_1) d\theta_1}{\int_0^\infty \exp\left\{-\frac{n(\bar{x}-\theta_1)^2}{2}\right\} \pi_1(\theta_1) d\theta_1},$$

where  $N(\bar{x}, n) = \int_{-\infty}^0 \exp\left\{-\frac{n(\bar{x}-\theta_2)^2}{2}\right\} \Phi\left(\frac{\theta_2}{\sqrt{2}}\right) d\theta_2$ . In Moreno, Bertolino and Racugno (1966) it was shown that  $\sup_{\pi_1 \in \Gamma_1} B_{21}(\bar{x}, n)$  is infinity for any

sample point. Since for  $\bar{x} \geq 0$  we have that  $\inf_{\pi_1 \in \Gamma_1} B_{21}(\bar{x}, n)$  is less than 1 we conclude that the Bayes factor is not robust with respect to the intrinsic priors.

A limiting procedure was In Moreno, Bertolino and Racugno (1966) introduced to overcome the lack of robustness. This procedure is based on the fact that when the prior for the simple model  $M_1$  is proper, then the intrinsic prior for the complex model always exists and it is unique.

The idea is to take the restriction of  $\pi_1^N(\theta_1)$  on an increasing sequence  $\{C_n\}$  of subsets of  $\Theta_1$  such that  $\int_{C_n} \pi_1^N(\theta_1) d\theta_1 < \infty$  and  $\lim_{n \rightarrow \infty} C_n = \Theta_1$ . Then we construct the associated sequence of intrinsic priors for  $M_2$  and we take the limit of the corresponding sequence of Bayes factors. Under rather general conditions the limit was proved to be independent on the particular sequence  $\{C_n\}$  we have chosen.

Applying this procedure to our example the limiting Bayes factor turns out to be

$$B_{21}(\bar{x}, n) = \frac{\Phi(\frac{\bar{x}\sqrt{n}}{\sqrt{2n+1}})}{\Phi(\bar{x}\sqrt{n})}.$$

On the other hand, the fractional priors for this problems are the following. For sequences  $\{b_n\}$  of the form  $b'_n = \frac{\sqrt{m_0}}{n}$ , and  $b''_n = \frac{\log n}{n}$ ,

$$\begin{aligned} F_{12}^{M_2}(\theta_2) &= \lim_{n \rightarrow \infty} [P_{\theta_2}] \frac{\int_{\Theta_1} f_1(z|\theta_1)^{b_n} \pi_1^N(\theta_1) d\theta_1}{\int_{\Theta_2} f_2(z|\theta_2)^{b_n} \pi_2^N(\theta_2) d\theta_2} \\ &= 1_{[0, \infty)}(\theta_2), \end{aligned}$$

so that

$$\pi_2(\theta_2) = \pi_1(\theta_2) 1_{[0, \infty)}(\theta_2),$$

where  $\pi_1(\theta_1)$  is any probability density on  $[0, \infty)$ . Therefore, the Bayes factor for any data  $z$  and any fractional prior turns out to be

$$\tilde{B}_{21}(z) = 1.$$

Therefore, the above sequences  $\{b'_n\}$ ,  $\{b''_n\}$  produce *fractional priors* that gives a non sensible Bayes factor.

If we choose  $b_n = \frac{m_0}{n}$  with  $m_0 = 1$ , the minimal training sample size, the class of fractional priors turns out to be

$$\pi_2(\theta_2) = \frac{k}{\sqrt{2\pi}} 1_{(-\infty, 0)}(\theta_2) + \Phi(\theta_2) \pi_1(\theta_2) 1_{(0, \infty)}(\theta_2)$$

where  $\pi_1$  is any prior in the class

$$\Gamma_2 = \{\pi_1 : \int_0^\infty \Phi(\theta_1) \pi_1(\theta_1) d\theta_1 = 1 - \frac{k}{\sqrt{2\pi}}, 0 \leq k \leq \sqrt{\frac{\pi}{2}}\},$$

$k$  being the value of  $\pi_1(\theta_1)$  at the discontinuity point  $\theta_1 = 0$ . The Bayes factor for  $\pi_1 \in \Gamma_2$  is given as

$$\tilde{B}_{21}(\bar{x}, n) = \frac{k \int_{-\infty}^0 \exp(-\frac{n(\bar{x}-\theta_2)^2}{2}) \Phi(\theta_2) d\theta_2 + \int_0^{\infty} \exp(-\frac{n(\bar{x}-\theta_1)^2}{2}) \Phi(\theta_1) \pi_1(\theta_1) d\theta_1}{\int_0^{\infty} \exp(-\frac{n(\bar{x}-\theta_1)^2}{2}) \pi_1(\theta_1) d\theta_1}.$$

It can be seen that  $\tilde{B}_{21}(\bar{x}, n)$  is not robust as the prior ranges over class  $\Gamma_2$ . To overcome this lack of robustness we can apply the same limiting procedure considered for the Bayes factor for intrinsic priors. A difficulty we find here is that it is not necessarily true that for a proper prior for the simple model and improper for the complex, the corresponding fractional prior for the complex model is proper. Indeed, if  $\pi_1(\theta_1)$  is a proper prior for  $M_1$  and  $\pi_2^N(\theta_2)$  the improper prior for  $M_2$ , the corresponding fractional prior for  $M_2$  is given as

$$\pi_2(\theta_2) = \pi_2^N(\theta_2) \lim_{n \rightarrow \infty} [P_{\theta_2}] \frac{\int_{\Theta_1} f_1(z|\theta_1)^{b_n} \pi_1(\theta_1) d\theta_1}{\int_{\Theta_2} f_2(z|\theta_2)^{b_n} \pi_2^N(\theta_2) d\theta_2},$$

which is not necessarily a probability density.

Fortunately, if in this example we take as  $\pi_1(\theta_1) = \frac{1}{b-a} 1_{(a,b)}(\theta_1)$ , the restriction of  $\pi_1^N(\theta_1)$  to the interval  $(a, b)$ , and  $b_n = \frac{1}{n}$  then the corresponding fractional prior for model  $M_2$  is

$$\pi_2(\theta_2) = \frac{1}{b-a} (\Phi(b-\theta_2) - \Phi(a-\theta_2)),$$

which is a probability density for any values of  $a$  and  $b$ . The Bayes factor for this fractional priors is

$$\tilde{B}_{21}^{(a,b)}(\bar{x}, n) = \frac{\int_{-\infty}^{\infty} \exp(-\frac{n(\bar{x}-\theta_2)^2}{2}) (\Phi(b-\theta_2) - \Phi(a-\theta_2)) \frac{1}{b-a} d\theta_2}{\int_0^{\infty} \exp(-\frac{n(\bar{x}-\theta_1)^2}{2}) \frac{1}{b-a} d\theta_1}.$$

The limit when  $a \rightarrow 0$  and  $b \rightarrow \infty$  results

$$\tilde{B}_{21}(\bar{x}, n) = \lim_{n \rightarrow \infty} \tilde{B}_{21}^{(a,b)}(z) = \frac{\Phi(\frac{\bar{x}\sqrt{n}}{\sqrt{n+1}})}{\Phi(\bar{x}\sqrt{n})},$$

which is very close to the limiting Bayes factor for intrinsic priors.

We have already said that the intrinsic methodology always generates a class of proper priors distributions (intrinsic priors) for model selection when the models are nested. The fractional methodology, however, not



necessarily generates proper priors. The following example illustrates this assertion.

**Example 3 *Two sided testing.*** Suppose we have to choose between the following two nested models,

$$M_1 : f_1(x|\theta_1) = N(x|0, \sigma_1^2), \quad \pi_1^N(\theta_1) \propto \frac{1}{\sigma_1} 1_{(0, \infty)}(\sigma_1),$$

$$M_2 : f_2(x|\theta_2) = N(x|\mu, \sigma_2^2), \quad \pi_2^N(\theta_2) \propto \frac{1}{\sigma_2^2} 1_{R \times (0, \infty)}(\mu, \sigma_2).$$

The parameter spaces are respectively  $\Theta_1 = 0 \times R^+$  and  $\Theta_2 = R \times R^+$  and the improper priors are given by the Jeffreys rule. It is easy to see that  $\psi_1(\theta_2) = \sqrt{\mu^2 + \sigma_2^2}$ .

For the data  $z = (x_1, x_2, \dots, x_n)$  and a given sequence  $\{b_n\}$ , some algebra shows that

$$\begin{aligned} F_{12}^{M_2}(\theta_2) &= \lim_{n \rightarrow \infty} [P_{\theta_2}] \frac{\int_{\Theta_1} f_1(z|\theta_1)^{b_n} \pi_1^N(\theta_1) d\theta_1}{\int_{\Theta_1} f_2(z|\theta_2)^{b_n} \pi_2^N(\theta_2) d\theta_2} \\ &= \lim_{n \rightarrow \infty} [P_{\theta_2}] \frac{\sqrt{n b_n}}{\sqrt{2\pi}} \left( \frac{s^2}{s^2 + \bar{x}^2} \right)^{b_n \frac{n}{2}}, \end{aligned}$$

where  $s^2 = \frac{1}{n} \sum_1^n (x_i - \bar{x})^2$  and  $\bar{x} = \frac{1}{n} \sum_1^n x_i$ . For sequences of the form  $\{b'_n\}$ ,  $\{b''_n\}$ , Assumption (A) is satisfied, but  $F_{12}^{M_2}(\theta_2) = 0$  and consequently no proper prior is fractional.

For sequences of the form  $\{b_n = \frac{m_0}{n}, m_0 = 1, 2, \dots, n-1\}$ ,

$$F_{12}^{M_2}(\theta_2) = \frac{\sqrt{m_0}}{\sqrt{2\pi}} \left( \frac{\sigma_2^2}{\sigma_2^2 + \mu^2} \right)^{m_0/2},$$

and thus the class of *fractional priors* is  $(\pi_1(\theta_1), \pi_2(\theta_2))$ , where

$$\pi_2(\theta_2) = \frac{\sqrt{m_0}}{\sqrt{2\pi}} \frac{\sigma_2^{m_0-2}}{(\sigma_2^2 + \mu^2)^{\frac{m_0-1}{2}}} \pi_1(\psi_1(\theta_2)),$$

and  $\pi_1(\theta_1)$  is any prior in the class

$$\begin{aligned} \Gamma_2 &= \{\pi_1(\theta_1) : \int_0^\infty \pi_1(\sigma_1) d\sigma_1 \\ &= 1, \int_0^\infty \pi_1(\sigma_1) d\sigma_1 \frac{\sqrt{m_0}}{\sqrt{2\pi}} \int_0^\pi (\sin \varphi)^{m_0-2} d\varphi = 1\}. \end{aligned}$$

It is easy to see that this class is empty for any  $m_0 \geq 1$ . In particular for  $m_0 = 2$ , the minimal training sample size for this problem, we have

$$\Gamma_2 = \{\pi_1(\theta_1) : \int_0^\infty \pi_1(\sigma_1) d\sigma_1 = 1, \int_0^\infty \pi_1(\sigma_1) d\sigma_1 = \frac{1}{\sqrt{\pi}}\},$$

that is clearly empty.

Therefore, the sequences  $\{b_n\}$  suggested in the literature of fractional methodology do not generate proper prior distributions for this two sided testing problem.

However, the class of intrinsic priors is

$$\Gamma_1 = \{\pi_1(\theta_1) : \int_{\Theta_1} \pi_1(\theta_1) d\theta_1 = 1, \int_{\Theta_1} \frac{1}{\sigma_1} \pi_1(\theta_1) d\theta_1 = 2(\sqrt{\pi} - 1)\},$$

that is not empty.

## 4 Conclusions

In this paper we have considered the intrinsic prior distributions (Berger and Pericchi, 1995 and 1996), and introduced the notion of fractional priors. This permits to focus the intrinsic and fractional methodologies as tools for generating proper prior distributions for model comparison from the conventional improper priors for estimation. The considered models have assumed to be nested and the main conclusions are:

The intrinsic priors always exist and form a class given by generalized moment constraints. The Bayes factor for intrinsic priors is not necessarily robust, but the limit intrinsic procedure (Moreno, Bertolino and Racugno, 1996) solves this lack of robustness.

The fractional methodology, however, not always generates proper priors (fractional priors). When fractional priors there exist, we have found that the sequence  $\{b_n = \frac{m_0}{n}, n \geq 1\}$  with  $m_0$  equal to the minimal training sample size, is the appropriate selection among those recommended by O'Hagan (1995). The associated Bayes factor is then very close to the Bayes factor for intrinsic priors. Furthermore, calculations are quite simple.

Extension of this theory to non-nested models is an interesting topic that deserves more research. It is a work in progress that will be formalized elsewhere.

## References

- [1] Akaike, H. (1983). Information measures and model selection. *Bull. Int. Statist. Inst.* **50**, 277-290.

- [2] Berger, J.O. and Mortera, J. (1995). Discussion to A. O'Hagan (1995). *J. R. Statist. Soc. B* **57**, 99-138.
- [3] Berger, J.O. and Pericchi, L.R. (1995). The intrinsic Bayes factor for linear models (with discussion). In *Bayesian Statistic*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Vol. 4, pp. 35-60. Oxford University Press.
- [4] Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction . *J. Am. Statist. Assoc.* **91**, 109-122.
- [5] Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press.
- [6] Kass, R.E. and Raftery, A. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773-795.
- [7] Kemperman, J.H.B. (1987). Geometry of the moment problem. *Proc. of Symposia in Applied Mathematics* **37**, 16-53.
- [8] Lavine, M., Wasserman, L. and Wolpert, R. (1993). Linearization of Bayesian robustness problems. *J. Am. Statist. Assoc.* **86**, 964-971.
- [9] Liseo, B., Moreno, E. and Salinetti, G. (1996). Bayesian robustness for classes of priors with given marginals (with discussion). *Proc. of the Workshop on Bayesian Robustness*, Eds. Berger, J.O., Betro, B., Moreno, E., Pericchi, L.R., Ruggeri, F., Salinetti, G. and Wasserman, IMS Lectures Notes-Monograph Series, Vol. 29, pp. 103-120.
- [10] Moreno, E., Bertolino F. and Ragcugno W. (1996). The intrinsic priors in model selection and hypotheses testing. Technical Report, University of Granada.
- [11] O'Hagan, T. (1995). Fractional Bayes factor for model comparison (with discussion). *J. R. Statist. Soc. B* **57**, 99-138.
- [12] Salinetti, G. (1994). Discussion on "An overview of Robust Bayesian Analysis" by J.O. Berger. *Test* **3**, 5-125.
- [13] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.
- [14] Spiegelhalter, D.J. and Smith, A.F.M. (1982). Bayes factor for linear and log-linear models with vague prior information. *J. R. Statist. Soc. B* **44**, 377-387

# A comparison of procedures based on inverse regression

**Simon J. Sheather**

*University of New South Wales, Australia*

**Joseph W. McKean**

*Western Michigan University, USA*

*Abstract:* Sliced inverse regression (SIR) was introduced by Li (1991) and Duan and Li (1991) as a dimension reduction technique that determines the number of linear combinations of the predictor variables needed to obtain a parsimonious regression model. It is well known that SIR is not robust to the effects of outliers nor can it always detect symmetric dependence. In this paper, we briefly outline another technique based on inverse regression which potentially overcomes these shortcomings of SIR in an important special case. Finally, we compare the effectiveness of the new technique with that of SIR on some real data sets.

*Key words:* Dimension reduction, sliced inverse regression, robustness.

AMS subject classification: 62J20, 62G35.

## 1 Introduction

Regression analysis is arguably one of the most widely used statistical techniques. A regression model expresses the mean of a response variable  $y$  as a function,  $f$ , of an explanatory variable  $\mathbf{x}$ , a  $p$ -dimensional column vector. Traditional parametric regression methods assume that the functional relationship between  $y$  and  $\mathbf{x}$  is known apart from some parameters, which must be estimated. When the assumed functional form is correct, a variety of methods (including least squares and robust methods) can be used to estimate the unknown parameters. However, in many applications any parametric model is at best an approximation to the true one and the search

for an adequate model becomes increasingly difficult as  $p$ , the number of predictors, increases.

Alternatively, nonparametric regression methods estimate the regression function without assuming a particular functional form. Recently, Wand and Jones (1995), Fan and Gijbels (1996) and Simonoff (1996) have provided comprehensive accounts of this field. Many nonparametric regression methods are based on the notion of local smoothing, that is, the estimate of  $f$  at any point of interest is based on a smoothed version of  $y$  in that region. Thus, the success of local smoothing depends on the existence of sufficiently many data points around each point of interest in the design space to provide adequate information about  $f$ . As the dimension of  $\mathbf{x}$  increases, larger and larger sample sizes are needed in order to ensure that there are sufficient data points around each point of interest. This problem has been appropriately referred to as the curse of dimensionality (Bellman, 1961). Hastie and Tibshirani (1990, pp. 83, 84) provide a simple, yet effective, example of this phenomenon. A number of approaches have been proposed to cope with the curse of dimensionality. Additive models (see Hastie and Tibshirani, 1990, Chapter 4) approximate  $f$  as the sum of nonparametric univariate functions of each of the  $p$  predictors. Alternatively, sliced inverse regression (Li, 1991 and Duan and Li, 1991) is a dimension reduction technique that does not rely on a complicated model-fitting process. Sliced inverse regression (*SIR*) determines the number of linear combinations of the  $p$  predictors needed to obtain a parsimonious model for  $f$ .

In the next section, we briefly outline a dimension reduction technique based on inverse regression. Finally, in Section 3 we compare the effectiveness of this new technique with that of *SIR* on some real data sets. We show that the new technique potentially overcomes two of the shortcomings of *SIR*, namely, a lack of robustness to outliers and a failure to detect some forms of symmetric dependence.

## 2 Inverse regression methods for dimension reduction

Consider the following general regression model

$$y = f(\beta'_1 \mathbf{x}, \dots, \beta'_k \mathbf{x}, \epsilon), \quad (1)$$

where  $f$  is an unknown arbitrary function,  $\mathbf{x}$  is a  $p$ -dimensional vector of predictors, and  $\epsilon$  is an  $n \times 1$  vector of errors which is assumed to be independent of  $\mathbf{x}$ . The integer  $k(< p)$  is the number of linear combinations of the predictors that are needed to summarize the dependence of  $y$  on  $\mathbf{x}$ . Li (1991, 1992), Cook and Weisberg (1991) and Schott (1994) provide

methods for determining the value of  $k$  based on sliced inverse regression.

The simplest nontrivial case occurs when  $k = 1$ , since then  $\beta_1' \mathbf{x}$  contains all the information from  $\mathbf{x}$  about  $y$ . If  $k \geq 2$  then commonly the aim is to reduce  $k$ . One possibility, for example, is to seek extra explanatory variables. These extra variables could include interaction terms, polynomial terms and/or dummy variables. Alternatively, the current set of explanatory variables could be changed using transformations. Further discussion of this issue can be found in Cook and Weisberg (1994, Chapter 8).

Thus arguably, an important special case is the decision as to whether  $k = 1$  or  $k > 1$ . In this paper, we focus on this problem.

Under the assumptions that model (1) holds with  $k = 1$  and that the distribution of  $\mathbf{x}$  is elliptically symmetric, Duan and Li (1991) obtained the following result

$$E(x_j|y) = E(x_j) + \gamma_j \kappa(y) \quad j = 1, \dots, p. \quad (2)$$

This result means that for each predictor  $x_j$ , the inverse regression function  $E(x_j|y)$  equals the mean of  $x_j$  plus some unknown function  $\kappa(y)$  times the constant  $\gamma_j$ . A crucial aspect of this result is that  $\kappa(y)$  does not depend on  $j$ . Hence, a graphical procedure to decide whether  $k = 1$  or not is to examine the  $p$  plots with  $x_j$  on the vertical axis and  $y$  on the horizontal axis to see if each plot has the same shape. Such a procedure is advocated by Cook and Weisberg (1994, Chapter 8).

Sheather and McKean (1997) have developed two nonparametric methods for testing whether  $k = 1$  or not. The two procedures are based on the following observation. Suppose that model (1) holds with  $k = 1$  and that the distribution of  $\mathbf{x}$  is elliptically symmetric. Then, for  $1 \leq i, j \leq p$

$$\log \left[ \left| \frac{E(x_j|y) - E(x_j)}{E(x_i|y) - E(x_i)} \right| \right] = \log \left[ \left| \frac{\gamma_j}{\gamma_i} \right| \right], \quad (3)$$

which is independent of  $y$ . In practice, the left side of (3) can be replaced by

$$L_{i,j} = \log \left[ \left| \frac{x_j - \bar{x}_j}{x_i - \bar{x}_i} \right| \right].$$

Thus, a test of  $k = 1$  against the alternative  $k > 1$  can be obtained by testing for each combination of  $i$  and  $j$  whether  $L_{i,j}$  is independent of  $y$ . Sheather and McKean (1997) have developed the following two tests.

- *Test1*: This test is based on dividing up each plot of  $L_{i,j}$  versus  $y$  into four quadrants and counting up the number of points in each quadrant. The quadrants are obtained by splitting both  $L_{i,j}$  and  $y$  into two groups depending on whether they are larger or smaller than their respective medians. If for a given  $i$  and  $j$ ,  $L_{i,j}$  does not depend on  $y$  then we expect

$n/4$  points to fall in each of the four quadrants. Departures from these expected frequencies are tested using a  $\chi^2$  goodness-of-fit test. Sheather and McKean (1997) have found that a significant result on *Test1* may indicate that polynomial terms are missing from the regression model.

- *Test2*: This test of independence is based on a runs statistic. In this case, the statistic used is the number of runs above and below the median value of  $L_{i,j}$  in each plot of  $L_{i,j}$  versus  $y$ . Sheather and McKean (1997) have found that a significant result on *Test2* may indicate that interaction terms are missing from the regression model.

For a more detailed description and discussion of *Test1* and *Test2* see Sheather and McKean (1997).

### 3 Examples

In this section, we discuss several examples involving real data taken from Cook and Weisberg (1994). In each example, the aim is to decide whether  $k = 1$  or  $k > 1$  in (1), that is, whether one linear combination of the predictors can adequately summarize the dependence of  $y$  on  $\mathbf{x}$ . The examples have been chosen to illustrate that the inverse regression technique potentially overcomes two of the shortcomings of *SIR*, namely, a lack of robustness to outliers and a failure to detect some forms of symmetric dependence.

The *R* – *code* software supplied with Cook and Weisberg (1994) was used to calculate *SIR*. In each example the default *R* – *code* settings for *SIR* were used. Rank based regression estimates were calculated using the experimental MINITAB command *rregress*. Once again all the default settings were used.

#### Example 1 *Ethanol Data*

The data consist of 87 observations obtained from an industrial experiment involving a one-cylinder engine using ethanol as a fuel. The response *NOx* is a measure of nitric oxide concentration in exhaust emissions. There are two predictors, *E* and *C*. *E* is the equivalence ratio, a measure of the fuel/air mixture while *C* is the compression ratio.

| <i>SIR</i> | <i>Test1</i> | <i>Test2</i> |
|------------|--------------|--------------|
| 0.345      | <0.0001      | 0.0003       |

Table 1:  $p$ -values for testing  $k = 1$  against the alternative  $k > 1$  for model (1).

Table 1 summarises the results from *SIR* as well as those from *Test1* and *Test2*. As discussed in Cook and Weisberg (1994, pp. 125-127), *SIR* has failed to detect the symmetric quadratic dependence of  $E$  on  $NOx$ , which is obvious in plots. On the other hand, *Test1* and *Test2* find very strong evidence that more than one linear combination of  $E$  and  $C$  is needed to adequately model  $NOx$ . In addition, these tests indicate that terms like higher order polynomial terms and interaction terms may be missing from the model.

The following regression model was fit to the data using least squares,

$$NOx = \gamma_0 + \gamma_1 C + \gamma_2 E + \gamma_3 C^2 + \gamma_4 E^2 + \gamma_5 C \cdot E + \epsilon. \quad (4)$$

Table 2 summarises the results. The interaction between  $C$  and  $E$  is highly significant as is the quadratic term in  $E$ . The adjusted  $R^2$  value for model (4) is 84.2% while the corresponding figure is 0.0% when the model without the quadratic and interaction terms is fit.

| Parameter  | Estimate | <i>t</i> -ratio | <i>p</i> -value |
|------------|----------|-----------------|-----------------|
| $\gamma_0$ | -24.26   | -15.702         | <0.001          |
| $\gamma_1$ | 0.22     | 1.876           | 0.864           |
| $\gamma_2$ | 56.76    | 20.592          | <0.001          |
| $\gamma_3$ | 0.00     | 0.565           | 0.574           |
| $\gamma_4$ | -29.79   | -21.465         | <0.001          |
| $\gamma_5$ | -0.24    | -3.868          | <0.001          |

Table 2: Least squares fit of model (4).

### Example 2 Australian Institute of Sport Data

The data were obtained from 102 male and 100 female athletes at the Australian Institute of Sport. Interest centers on modeling *LBM* (lean body mass) as a function of *Ht* (height in centimetres), *Wt* (weight in kilograms) and *RCC* (red cell count). Following Cook and Weisberg (1994, pp. 122-125) we shall analyse the data for the female and male athletes separately.

#### Female Athletes

Table 3 summarises the results from *SIR* as well as those from *Test1* and *Test2* for the data on the 100 female athletes. The *p*-values reported for *Test1* and *Test2* are the minimum of those obtained from 3 pairwise tests. The row headed '1 point removed' gives the results when the case corresponding to the largest value of *LBM* is removed from the data, while the



row headed ‘5 points removed’ gives the results when the 5 cases marked with a  $\times$  in Figure 8.5 of Cook and Weisberg (1994, p. 125) are removed from the data.

|                  | <i>SIR</i> | <i>Test1</i> | <i>Test2</i> |
|------------------|------------|--------------|--------------|
| All data         | 0.026      | 0.686        | 0.005        |
| 1 point removed  | 0.229      | 0.836        | 0.008        |
| 5 points removed | 0.617      | 0.724        | 0.007        |

Table 3:  $p$ -values for testing  $k = 1$  against the alternative  $k > 1$  for model (1).

As discussed in Cook and Weisberg (1994, pp. 124-125), *SIR* is not robust to the effects of outliers. In this case removing just one of the 100 data points produces a 10 fold change in  $p$ -value obtained from *SIR*. On the other hand, *Test1* and *Test2* change little when a small number of data points are removed from the data. In addition, *Test2* finds strong evidence that more than one linear combination of *Ht*, *Wt* and *RCC* is needed to adequately model *LBM*. In addition, this test indicates that interaction terms may be missing from the model.

The following regression model was fit to the data using rank based regression,

$$LBM = \gamma_0 + \gamma_1 Ht + \gamma_2 Wt + \gamma_3 RCC + \gamma_4 Ht \cdot Wt + \gamma_5 Ht \cdot RCC + \gamma_6 Wt \cdot RCC + \epsilon. \tag{5}$$

Table 4 summarises the results when model (5) is fit to all 100 data points. In this case, the interaction between *Ht* and *Wt* is highly significant. When the 5 points referred to in Table 3 are removed and model (5) is refit, the  $p$ -value for the interaction between *Ht* and *Wt* increases to 0.056. Thus, some but not all of the significance of this interaction term is due to these 5 points.

| Parameter  | Estimate | $t$ -ratio | $p$ -value |
|------------|----------|------------|------------|
| $\gamma_0$ | -61.400  | -2.567     | 0.012      |
| $\gamma_1$ | 0.719    | 2.958      | 0.004      |
| $\gamma_2$ | 0.804    | 1.529      | 0.130      |
| $\gamma_3$ | -0.007   | -0.134     | 0.894      |
| $\gamma_4$ | -0.006   | -2.744     | 0.007      |
| $\gamma_5$ | -0.056   | -1.518     | 0.132      |
| $\gamma_6$ | 0.169    | 1.754      | 0.082      |

Table 4: Rank based regression fit of model (5) to all 100 female athletes.

*Male Athletes*

Table 5 summarises the results from *SIR* as well as those from *Test1* and *Test2* for the data on the 102 male athletes. The  $p$ -values reported for *Test1* and *Test2* are the minimum of those obtained from 3 pairwise tests. The row headed ‘2 points removed’ gives the results when the cases corresponding to the two largest values of *LBM* are removed from the data.

|                  | <i>SIR</i> | <i>Test1</i> | <i>Test2</i> |
|------------------|------------|--------------|--------------|
| All data         | 0.017      | 0.950        | 0.111        |
| 2 points removed | 0.173      | 0.831        | 0.108        |

Table 5:  $p$ -values for testing  $k = 1$  against the alternative  $k > 1$  for model (1).

These data also illustrate that *SIR* is not robust to the effects of outliers. In this case removing just two of the 102 data points produces a 10 fold change in  $p$ -value obtained from *SIR*. On the other hand, *Test1* and *Test2* change little when a small number of data points are removed from the data. In addition, neither *Test1* nor *Test2* finds strong evidence that more than one linear combination of *Ht*, *Wt* and *RCC* is needed to adequately model *LBM*.

The following regression model was fit to the data using rank based regression,

$$LBM = \gamma_0 + \gamma_1 Ht + \gamma_2 Wt + \gamma_3 RCC + \gamma_4 Ht \cdot Wt + \gamma_5 Ht \cdot RCC + \gamma_6 Wt \cdot RCC + \epsilon. \quad (6)$$

Table 6 summarises the results when model (6) is fit to all 102 data points. In this case, none of the interactions are significant. When the 2 points referred to in Table 5 are removed and model (6) is refit, once again none of the interaction terms are significant.

| Parameter  | Estimate | $t$ -ratio | $p$ -value |
|------------|----------|------------|------------|
| $\gamma_0$ | -69.870  | -1.072     | 0.286      |
| $\gamma_1$ | 0.590    | 1.344      | 0.182      |
| $\gamma_2$ | 0.716    | 1.903      | 0.060      |
| $\gamma_3$ | 6.650    | 0.544      | 0.588      |
| $\gamma_4$ | -0.002   | -1.070     | 0.287      |
| $\gamma_5$ | -0.062   | -0.749     | 0.228      |
| $\gamma_6$ | 0.058    | 0.997      | 0.321      |

Table 6: Rank based regression fit of model (6) to all 102 male athletes.

In summary, for the athletes data the nonparametric procedure of Sheather and McKean (1997) seems to correctly identify the case (i.e., the male ath-

letes) when a single linear combination of  $Ht$ ,  $Wt$  and  $RCC$  adequately models  $LBM$  as well as the case (i.e., the female athletes) when more than one linear combination is needed. On the other hand,  $SIR$  indicates for both the female and male athletes that more than one linear combination is needed when all the data are used while it indicates the opposite when a small number of outliers have been removed from the data.

## References

- [1] Cook, R. and Weisberg, S. (1991). Comment on Li (1991). *J. Am. Statist. Assoc.* **86**, 328-332.
- [2] Cook, R. D. , and Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.
- [3] Duan, N. and Li, K. C. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* **19**, 505-530.
- [4] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.
- [5] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman and Hall.
- [6] Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Assoc.* **86**, 316-342.
- [7] Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Assoc.* **87**, 1025-1039.
- [8] Schott, J. R. (1994). Determining the dimensionality for sliced inverse regression. *J. Am. Statist. Assoc.* **89**, 141-148.
- [9] Sheather, S. J. and McKean, J. W. (1997). A dimension reduction technique based on inverse regression. Unpublished manuscript.
- [10] Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer.
- [11] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

# Interactive graphical modeling methods for regression

Nate Wetzel

*State University of New York, Binghamton, USA*

*Abstract:* Identification of curvature in regression models is an important aspect of data analysis. Partial residual plots have played a major role. Recently a new class of plots has been developed. They are called CERES plots and include partial residual plots as a special case. Implementation of these plots necessitates modeling the relationships between certain covariates. If these relationships are linear, a partial residual plot is obtained. However, if the relationships are nonlinear, the more general CERES plot is obtained. Generalized additive models (GAM) are another method for identifying and estimating curvature. Again, implementation of a GAM requires modeling the relationships between covariates and the response. Here, we motivate and describe key features of interactive, graphical methods which construct CERES plots and/or GAMs.

*Key words:* Partial residual plots, CERES plots, Generalized Additive Models, XLISP-STAT, S-PLUS.

AMS subject classification: 62G07.

## 1 Introduction

Conditional expectation residual plots (CERES plots, see Cook, 1993) and Generalized Additive Models (GAMs, see Hastie and Tibshirani, 1990) have been developed in the literature as diagnostic and modeling tools for regression analysis. These methods are designed to detect curvilinear relationships between selected covariates and the response variate in regression. When used interactively, these methods can help detect outliers, give information about possible heteroscedasticity.

In this paper, we outline the basic theory and assumptions underlying CERES plots and GAMs. Using simulated data, we then illustrate how

these methods are used and implemented. Both of these methods rely on the use of scatterplot smoothers. The examples are intended to highlight the usefulness of an implementation which (1) shows the data and associated scatterplot smoothers and (2) has an interactive interface so that smoothers can be easily changed and results compared.

## 2 CERES plots and GAMs - a primer

Consider the regression model (given  $X_1$  and  $X_2$ )  $Y = \alpha_0 + g_1(X_1) + g_2(X_2) + \epsilon$ , where  $\alpha_0$  is an unknown constant and  $g_1$  and  $g_2$  are unknown functions with  $E(g_i(X_i)) = 0$  and  $E(\epsilon|X_1, X_2) = 0$ . In general,  $X_1$  and  $X_2$  may be random vectors, but for the purposes of this paper,  $X_1$  and  $X_2$  are random variables. In other words, for the purposes of this paper, there are two predictor variables.

The idea behind CERES plots (see Cook, 1993) is that if  $g_1$  is the identity function and  $E(X_1|X_2)$  is known, then a CERES plot will display the function  $g_2$ . This display will be with error and possible vertical shift. In practice,  $E(X_1|X_2)$  is unknown, so we estimate it by smoothing the plot of  $X_1$  versus  $X_2$ , and then estimate  $g_2$  by smoothing the CERES plot which is obtained by assuming that our estimate of  $E(X_1|X_2)$  is correct. An implementation of CERES plots using the XLISP-STAT software (see Tierney, 1990) is given in a paper by Wetzel (1996).

GAMs have the additional assumption that  $\epsilon$  is independent of  $(X_1, X_2)$ , and the basic idea is that if we know  $g_1$  then  $E(Y - \alpha_0 - g_1(X_1)|X_2) = g_2(X_2)$ . In practice  $g_1$  is unknown, so we use an iterative algorithm to estimate  $g_1$ , then  $g_2$ , then  $g_1$ , etc. An implementation of GAMs is given in the S-PLUS software.

The theory underlying both CERES plots and GAM is powerful; however, when used in practice, we need the implementations to be interactive enough so that we can be critical users. When using the above methods in exploratory data analysis, we need to be able to look 'behind the scenes' to, in the CERES case, see the smooth which estimates  $E(X_1|X_2)$ , and in the GAM case, see the iterative process. In order to critically use these methods, we must be able to adjust and see the new results quickly. This need is demonstrated in the next section.

## 3 The need for interactive methods

In this section, we will look at a few examples which illustrate the need for interactive methods when using either CERES plots and/or GAM.

### 3.1 The influence of the choice of smoother in GAMs

The first example involves randomly generated data with the following distributions: (1)  $X_1 \sim N(0, 1)$ , (2)  $X_2|X_1 \sim N(4 + .25 * X_1, .01)$ , (3)  $Y = 6 + X_1^2 + \log(X_2 - 3)$ .

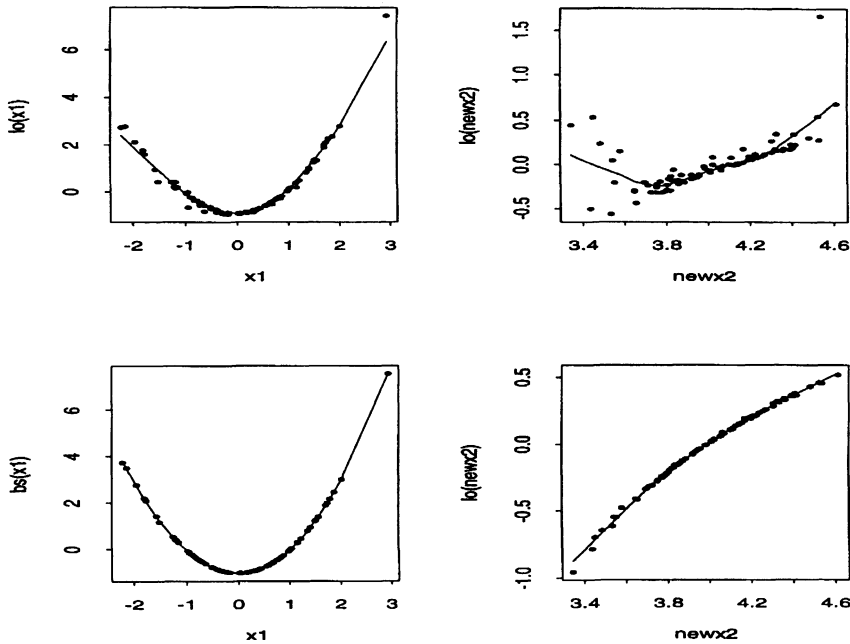


Figure 1: Estimated additive functions from GAM.

Two Generalized Additive Models were fitted and the summary plots from S-PLUS are given in Figure 1. The top two plots are from the fit given by the Splus command `gam(y ~ lo(x1) + lo(x2))` and the bottom two plots from the command `gam(y ~ bs(x1) + lo(x2) , bf.maxit=20)`. Here `lo` and `bs` correspond to a loess fit and b-spline fit, respectively.

The summary plots from S-PLUS show the estimated functions  $\hat{g}_1$  and  $\hat{g}_2$  as well as the points used to estimate these curves. The horizontal axes have the predictors  $X_1$  and  $X_2$ , and the vertical axes have what can be thought of as partial residuals. They are partial residuals, but are weighted in a non-trivial way. As analysts, we are to know that if the points are closer to the curve, then the fit will have smaller residual sum of squares.

Notice that the predicted function of  $X_1$  in both cases appears to be quadratic. In the first GAM, we see a loess curve, and in the second a b-spline fit. However, the predicted functions for  $X_2$  are very different. In the first case it appears that  $Y$  is dependent on  $X_2$  quadratically, and in the second, we see the true logarithmic relationship.

This shows that the choice of smoothers in GAM is very important. Ideally, a dedicated data analyst would see that in the first GAM, the loess smooth for  $X_1$  is under-fitting for both the small and large values of  $X_1$ . An interactive interface should allow them to interactively change the smoother for  $X_1$ . An interactive investigation of the 'outlier', when  $X_1 \approx 3$ , may also be informative. In this case, deleting the 'outlier' does not significantly change the predicted model, while changing the smoother does.

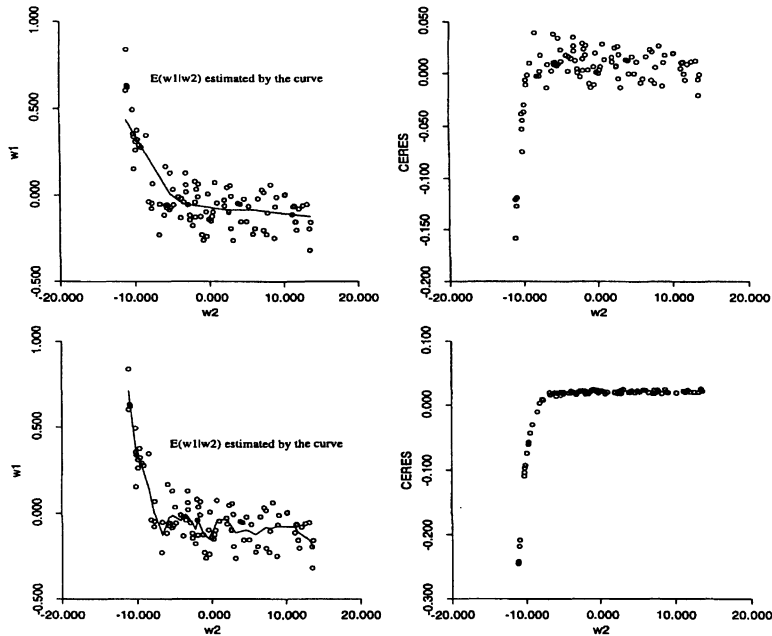


Figure 2: Estimated  $E(W_1|W_2)$  and associated CERES plot.

### 3.2 Looking at intermediate plots for CERES

As demonstrated in Wetzel (1996), the intermediate step of estimating  $E(X_1|X_2)$  has a great influence on the resulting prediction. Here, we generate data similar to Wetzel (1996): (1)  $W_2 \sim \text{Uniform}[1, 26]$ , (2)  $W_1|W_2 \sim N(1/W_2, .01)$ , (3)  $Z = W_1 + 1/(1 + \exp(-W_2))$ . In this example  $g_2(w_2) = 1/(1 + e^{-w_2})$ . A graph of the function  $g_2$  after being shifted both horizontally and vertically, looks exactly like the lower right plot in Figure

2. The plots in Figure 2 show two choices of smoother used to estimate  $E(W_1|W_2)$  and the associated CERES plot. These plots were obtained using the XLISP-STAT package and the code developed in Wetzel (1996). The leftmost plot shows the smooth used to estimate  $E(W_1|W_2)$  and the rightmost plot shows the CERES plot. ( $W_1$  and  $W_2$  are centered in all plots.) As analysts, the CERES plot should be smoothed and the resulting smoother used to estimate  $g_2$ . Again, the further the points in the CERES plot are from the smooth, the larger the residual sum of squares for a final fit.

The plots in Figure 2 show that the choice of smoother has a large effect of our perception of the amount of noise in the prediction of  $g_2(W_2)$ . Again, a dedicated data analyst would be able to interactively adjust the smoother and see that for a coarse smooth for  $E(W_1|W_2)$ , the apparent noise in the prediction of  $g_2(W_2)$  is reduced. Although experience has shown that a coarse smooth for  $E(W_1|W_2)$  often results in a more accurate display of  $g_2$ , the point here is that the user should easily be able to experiment with smoothers and parameters. In this case, experimentation allows us to *see* that it is not the coarseness of the smooth that makes the second CERES plot give a more accurate smooth, but instead it is the fact that the coarser smooth is closer to the truth for small values of  $W_2$ . In fact, if we estimate  $E(W_1|W_2)$  with a piecewise linear using only two lines, the CERES plot looks virtually the same as the the lower right plot of Figure 2.

### 3.3 Influential Points

Imagine that in our first example, we had an error in measurement in the observation where  $X_1$  is largest. This point is already a suspected outlier, but imagine that instead of a response value of 15.04, a response of 19.04 was recorded. We fit the same GAMs used in section 3.1, and the plots in Figure 3 are obtained.

The error in measurement actually serves to allow the loess smoother for  $X_1$  to begin to capture the true parabolic relationship between  $X_1$  and  $Y$  for positive values of  $X_1$ . GAM. Notice that although the estimated functions for the model fit by the Splus command `gam(y ~ lo(x1) + lo(x2))` do not differ much from those in Figure 1, the observations with values of  $X_2$  between 4.2 and 4.6 are fit much better when we have the error in measurement.

This shows that single points may be highly influential in the estimated fit as well as the perception of fit.

## 4 Interactive Methods

The above examples illustrate that there is a need for interactive meth-



ods which will aid the data analyst in understanding the regression. Such methods should:

- allow the user to interactively change the smoothers.
- allow the user to investigate the influence of individual points.
- not be too cumbersome for practical use.

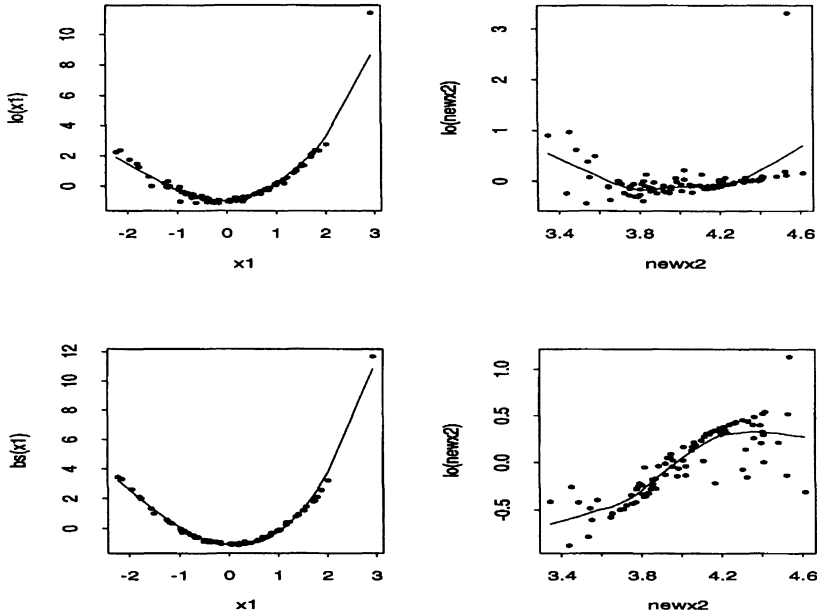


Figure 3: Estimated additive functions from GAM.

In Wetzel (1996), such a set of methods was developed for CERES plots. The estimation of  $E(X_1|X_2)$  is displayed and the user may easily change smoothing parameters, and/or the points used in the calculation. A set of reasonable defaults were established, but the analyst was presented with all of the relevant plots. Smoothers can be changed and points deleted and after a few mouse clicks, all of the plots are updated.

An outline of a similar set of methods for GAM is described below. Since the GAM procedure is iterative, we need to be able to monitor the process through all of the iterations. This process is described by the following algorithm (see Hastie and Tibshirani, 1990).

- initialize:  $g_1 = g_1^0$ ,  $g_2 = g_2^0$ .
- cycle:  $g_1^i =$  apply a smoother to the plot of  $Y - g_2^{i-1}(X_2)$  versus  $X_1$ ,  
 $g_2^i =$  apply a smoother to the plot of  $Y - g_1^i(X_1)$  versus  $X_2$ ,

- continue until  $g_1$  and  $g_2$  don't change

Interactive methods for GAM should allow the user to see each of the plots which are smoothed. An analyst's perception of the appropriateness of these smoothes will allow that analyst to proceed with another iteration. If the smooth is seen to be inappropriate, a different smoother is chosen, and then another iteration is performed. Similarly, if a point is deemed an outlier, the analyst will take appropriate action and continue with another iteration. The software should keep track of which smoothers were used at which iteration as well as which points were used. At some point, we need to 'continue until  $g_1$  and  $g_2$  don't change.' At this point, the observations which will be used and the smoothers should be fixed. We have not proven a result, but it seems clear that the first few steps of the iterative procedure should have little bearing on convergence results.

In order to illustrate these methods, we return to our first example. Using Xlisp-Stat and an initialization of  $g_1^0 = 0$  and  $g_2^0 = 0$ , we obtain the plots in Figure 4. The ordering of these plots is left to right, and top to bottom. The curves shown in the plots are the smoothes used to estimate  $g_1^i$  and  $g_2^i$ . The indicated numerical argument for lowess is the value used to call the `lowess` function in Xlisp-Stat. No weighting is used in this example. For example, the plot in the upper right hand corner shows  $Y - \bar{Y} - g_1^1(X_1)$  versus  $X_2$ , where the smooth shown in the upper left plot is used for  $g_1^1$ .

We notice that the lowess smooth used in the first iteration under estimates at the extremes, so in the second iteration we use a 2nd degree polynomial and obtain better estimates for both  $g_1$  and  $g_2$ .

The code for such methods is currently underway; the first plot in Figure 4 was created with a command from the keyboard, but the other five plots in Figure 4 were created with a series of mouse clicks. A final mouse click had the iteration continue until a crude convergence criteria was met. The final fits do not appear significantly different from the third row of plots in Figure 4.

## 5 Discussion

Finally, it should be clear that there is a connection between CERES plots and GAMs. Both find nonlinear relationships between the response and predictors. CERES plots assume that all of the predictors act linearly except for one. GAMs add additional assumptions to the errors. Berk and Booth (1995) compare CERES and GAMS to each other as well as other methods. Also, since at each stage in the iterative process GAMs use partial residual plots, and partial residual plots assume that the relationship between predictors is at most linear, strong nonlinear relation-

ships between the predictors may result in poor GAM performance. An approach to combining these two ideas is being investigated (see Croos-Dabrera, 1994). Implementation of these methods should allow interaction as described above.

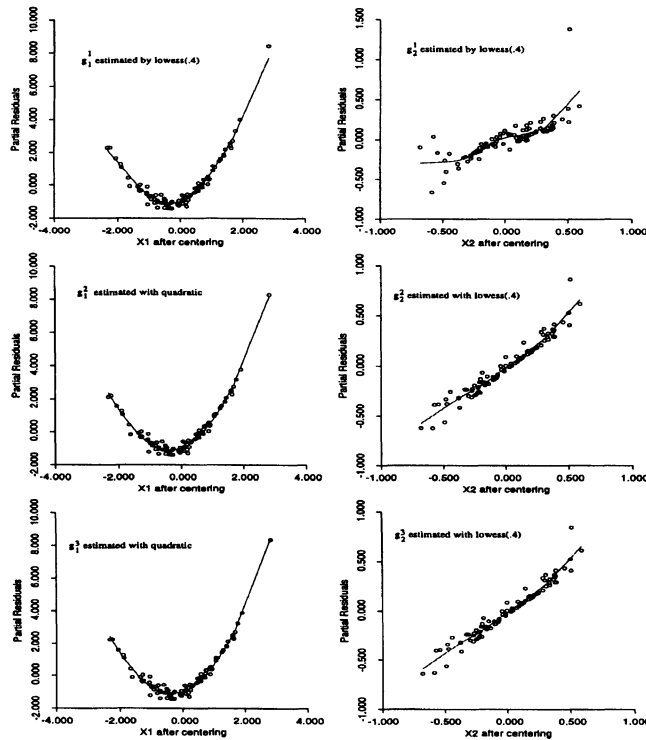


Figure 4: Sequential Partial Residual Plots.

## References

- [1] Berk, K. N. and Booth, D. E. (1995). Seeing a curve in multiple regression. *Technometrics* **37**, 385–398.
- [2] Cook, R. D. (1993). Exploring partial residual plots. *Technometrics* **35**, 351–362.
- [3] Croos-Dabrera, R. (1994). Graphical analysis of curvature in semiparametric generalized linear models. Ph.D. dissertation, University of Minnesota - School of Statistics.
- [4] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. New York: Chapman and Hall.
- [5] Tierney, L. (1990). *LISP-STAT*. New York: Wiley.
- [6] Wetzel, N. (1996). Graphical data modeling methods using CERES plots. *J. Statist. Comput. Simul.* **54**, 37–44.

# Nonparametric bounds for the probability of future prices based on option values

Gilbert W. Bassett Jr.

*University of Illinois, Chicago, USA*

*Abstract:* Interest in using option prices to estimate implied probabilities of stock values has emerged out of evidence suggesting the lognormal assumption of the Black Scholes model is no longer accurate. Most of the evidence relates to stock index option prices, especially since October 1987. The Black Scholes model assumes stock prices follow a geometric Brownian motion in continuous time - a lognormal distribution in discrete time. The standard deviation or volatility of the stock price process is the only unknown value in the formula so that implied standard deviations (volatilities) can be deduced from observed option prices. Prior to 1987, however, the implied volatility tended to curve upwards at far from at-the-money strike prices. Because of its shape, the relation came to be known as the "smile". The smile implies a fat-tailed underlying distribution, a long recognized feature of stock prices. Since the 1987 crash, the smile has deteriorated much farther from what it is supposed to look like under lognormality. Not flat and now not even a smile, it skews significantly to the left, indicating large probabilities of price decreases. This has led to recent proposals that focus on nonparametric estimates of the shape of the underlying distribution. A similar approach is followed here, but rather than estimating specific distributions, bounds are derived for the set of probability distributions that could have generated observed prices. These may be considered as either the first step toward identifying a single estimate, or as a nonparametric range of estimates for the underlying probabilities.

*Key words:* Option prices, nonparametric statistics, futures prices, robustness, Black Scholes model.

AMS subject classification: 90A09.

## 1 Introduction

Interest in using option prices to estimate implied probabilities of stock values has emerged out of evidence suggesting the lognormal assumption of the Black Scholes model is not very accurate. Most of the evidence relates to stock index option prices, especially since October 1987. The Black Scholes model assumes stock prices follow a geometric Brownian motion in continuous time—a lognormal distribution in discrete time. The standard deviation or volatility of the stock price process is the only unknown value in the formula so that implied standard deviations (volatilities) can be deduced from observed option prices.<sup>1</sup> Since European calls with different strike prices, but the same expiration date, are governed by the same probability distribution, they will have identical implied volatilities when the lognormal specification is valid. Prior to 1987, however, the implied volatility tended to curve upwards at far from at-the-money strike prices. Because of its shape, the relation came to be known as the "smile". The smile implies a fat-tailed underlying distribution, a long recognized feature of stock prices; see, e.g., Mandelbrot (1963) and Fama (1965). Since the 1987 crash, the smile has deteriorated much farther from what it is supposed to look like under lognormality. Not flat and now not even a smile, it skews significantly to the left, indicating large probabilities of price decreases - what Rubinstein (1994) calls, "crashophobia".

The initial response to understanding the smile was to generalize the geometric Brownian motion model by making volatility random, while maintaining lognormality. Stochastic volatility models generate smiles because they make the (unconditioned by volatility) underlying distribution fatter-tailed than lognormal. (This corresponds to the well known Monte Carlo trick for generating fat-tailed distributions: generate normal variates, but with different variances).

The recent evidence on implied volatilities has led to proposals that focus on estimating the entire shape of the underlying distribution; see, e.g., Shimko (1993) and Rubinstein (1994). These methods are nonparametric and do not presume lognormality. A similar approach will be pursued here, but rather than estimating specific distributions, bounds are derived for the set of probability distributions that could have generated observed prices. These may be considered as either the first step toward identifying a single estimate, or as a nonparametric range of estimates for the underlying probabilities.

---

<sup>1</sup>As the only free parameter, volatility stands for everything that affects option prices, but which is not in the model; see Figlewski (1989).

## 1.1 Convexity

The estimates are based on a connection between convex functions and cumulative probability distributions. To any cdf  $F$  there is the associated convex function:

$$g(K) = \int_{-\infty}^K F(s)ds.$$

Convexity of  $g$  follows from the fact that its derivative is the nondecreasing cdf  $F$ .<sup>2</sup> The resulting convex function is not arbitrary as its derivative must also satisfy the boundary conditions of a cdf;  $F(x) \rightarrow 1$  and  $0$  as  $x \rightarrow \pm\infty$ . Conversely, to a convex function  $g$  (satisfying the boundary conditions), there is the associated cdf that is its first derivative.<sup>3</sup>

This convexity correspondence arises with option valuation because arbitrage-free option prices are necessarily a convex function of strike prices. Hence there is always a probability distribution implicit in such option prices. Since convexity follows from arbitrage-free valuation alone, there exists an implied cdf given *any* specification of risk preferences and *any* stochastic process for the underlying stock price.

When investors are risk-neutral the implied probability distribution is identical to the cdf of the underlying stock price at expiration. This also occurs under assumptions, such as those in the Black Scholes model, where call values are determined independently of investor risk preferences. In the Black-Scholes model the underlying stock price is assumed to follow a geometric Brownian motion, a lognormal distribution in discrete time (or a binomial process that is Brownian motion in the limit). When prices follow such a process, call values are determined by arbitrage considerations alone - risk preferences do not matter - and the implied risk-neutral cdf is the same as the one that governs the stock at expiration.

The existence of implied probabilities however holds generally and does not require the lognormal specification. There is an implied distribution

---

<sup>2</sup>This does not require that  $g$  be differentiable or that  $F$  be continuous. When  $F$  corresponds to a discrete distribution,  $g$  is a polyhedral convex function.  $F$  can be recovered from  $g$  via the directional derivative, where the direction is determined by the left/right continuity convention adopted for cdfs. For properties of convex functions and their derivatives see Rockafellar (1970).

<sup>3</sup>This convexity/probability connection arises in unexpected places. One case is the generalized Lorenz curve used for determining second degree stochastic dominance. The generalized Lorenz curve is the  $g$  function derived from the quantile (inverse of  $F$ ) income distribution. A different context where the convexity is useful is in verifying that a particular linear function of regression quantiles defines an empirical cdf. It is not at all obvious, for example, that a combination of regression quantiles defines an empirical cdf until the combination is recognized as the derivative of a convex function; see Bassett and Koenker (1982, Theorem 2.1, p.409) and Koenker and Bassett (1978).

when  $F$  is not lognormal, and even if investors do not have explicit probability assessments about future values. In situations where risk matters, there will be an implied cdf, though it need not agree with the process that governs prices at expiration.

Section 2 briefly describes the connection between convex, arbitrage-free, call values and the implied risk-neutral probability distribution. Arbitrage-free call values were first described in Merton (1973) (also, see Cox and Rubinstein (1985)), and Breeden and Litzenberger (1979)). By proceeding from prices to inferred probabilities, we reverse the standard approach in which prices arise out of causally prior probabilities. Bounds for the underlying cdf, given a discrete set of option prices, are presented in Section 3 along with the modifications needed when convexity is invalidated by non-zero transaction costs, bid-ask spreads, and nonsynchronous prices.

## 2 Option values and probabilities

Let  $c(K)$  denote the time  $t$  value of a European call option with expiration  $T > t$ . The underlying asset's current value is  $S_t$  and the unknown value at expiration is  $S_T$ . Suppose initially that no dividends are paid between  $t$  and  $T$ , and that there is a continuum of strike prices in the interval  $[0, K_{\max}]$ , where  $K_{\max}$  is large enough that  $c(K_{\max}) = 0$ .

### 2.1 Risk-neutral call values

Let the risk free rate of return be denoted by  $r$ . Suppose investors are risk-neutral; that is, indifferent between a riskless return and a random return with the same expected value; see Harrison and Pliska (1981). Prices in equilibrium are then determined by expected values. Let  $F(s) = \Pr[S_T \leq s]$  represent a cdf for  $S_T$ . Since the value  $S_t$  invested today in risk free bonds at rate  $r$  yields  $e^{r(T-t)}S_t$  at time  $T$ , the expected value for a risk-neutral investor must grow at the risk free rate. Hence, the expectation of  $F$  is required to satisfy,  $E(S_T) = e^{r(T-t)}S_t$ , but otherwise  $F$  is arbitrary.

The value of  $c(K)$  at expiration is the random variable,  $\max\{0, S_T - K\}$ , whose expected value is,

$$E[\max\{0, S_T - K\}] = \int_K^\infty (s - K)dF(s).$$

Integration by parts yields the convenient expression,

$$E[\max\{0, S_T - K\}] = E(S_T) - \int_0^K (F(s) - 1)ds.$$

Finally, let  $g(K : F)$  denote the discounted present value of the expectation,

$$g(K : F) = e^{-r(T-t)} E[\max\{0, S_T - K\}] = S_t + e^{-r(T-t)} \int_0^K (F(s) - 1) ds \quad (1)$$

This provides the basic expression for determining prices from probabilities, or probabilities from prices. In a risk neutral world with distribution  $F$ , call prices are  $c(K) = g(K : F)$ . Alternatively, given call prices  $c(K)$  there exists the implied distribution  $F$  such that  $g(K : F) = c(K)$ .

From expression (1) we see that the first derivative of  $g$  with respect to  $K$  recovers the underlying cdf, and the second derivative produces the probability density; the respective derivatives are  $F(K) - 1$  and  $f(K)$ , each scaled by the discount factor. Since the functions are related by integration/differentiation, the call price curve will be smoother than the cdf, which will be in turn smoother than the density. Finally, expression (1) shows that  $g(0 : F) = S_t$ , reflecting the equivalence between the underlying stock and a zero-strike call option. Since the shares, but not the options, may receive dividends, the identity has to be modified when dividends are nonzero.

In view of (1), risk neutral call prices satisfy certain basic properties: there has to be an  $F$  such that  $c(K) = g(K : F)$ . What does this imply about the form of  $c(K)$ ? The following are the features of risk-neutral call values.

1.  $c(K)$  is nonnegative with  $c(0) = S_t$ .
2.  $c(K)$  is decreasing with  $-e^{r(t-T)} < dc/dK < 0$ .
3.  $c(K)$  is convex.

The first property follows from  $\max\{0, S_T - K\} \geq 0$ ; the second says the derivative is nonpositive,

$$dc/dK = dg/dK = e^{-r(T-t)} (F(K) - 1) \leq 0;$$

and the third follows from the fact that the first derivative increases with  $K$ , or, when there is a density the second derivative is  $e^{-r(T-t)} f(K) \geq 0$ .

## 2.2 Arbitrage-free call prices

Suppose now that risk neutrality is relaxed and investors have arbitrary risk preferences and perhaps even know nothing of probability. Suppose, however, that all arbitrage opportunities are exploited; that is, call prices  $c(K)$  are such that there are no riskless profit opportunities from buying or selling calls, or investing at the risk free interest rate (assuming zero transaction costs). What does this imply about  $c(K)$ ?



It is now well known that arbitrage-free call prices are identical to risk-neutral call prices; call values are arbitrage-free if and only if there is an  $F$  such that  $c(K) = g(K : F)$ . In a risk neutral world the  $F$  representing the beliefs of investors is the same as the  $F$  implied by call prices, whereas in a non risk-neutral world the  $F$  implicit in call prices is the equivalent martingale measure. This may seem surprising since the arbitrage-free requirement says nothing about probabilities. The intuition behind the equivalence is similar to Dutch book explanations for coherent beliefs regarding probability assessments. When your beliefs are not consistent with the probability axioms you can make book against yourself and win (lose). To see why arbitrage-free call values must be nonnegative, decreasing, and convex, as well as the risk free arbitrage opportunities that would occur if one of the conditions is violated, see Cox and Rubinstein (1985, p.237); also, see Cox and Ross (1976) for option valuation with stochastic processes other than geometric Brownian motion.

### 2.3 Call price curves

Given expression (1) we can identify an  $F$  from a  $c(K)$ , or a  $c(K)$  from an  $F$ . The types of curves that arise in simple special cases are illustrated in the following examples. For simplicity  $r$  is assumed to be zero.

#### 2.3.1 Discrete probabilities

Suppose  $S_T$  is a discrete random variable that takes values  $s_j$  with probabilities  $p_j$ ,  $j = 1, \dots, J$ . Then the cdf  $F(s) = \Pr[S_T \leq s]$  is a discontinuous jump function, and call prices are a linear spline,

$$c(K) = c(s_j) - (1 - F(s_j))(K - s_j), \quad s_j \leq K \leq s_{j+1}.$$

This situation is illustrated in Figure 1.

#### 2.3.2 Histogram probabilities

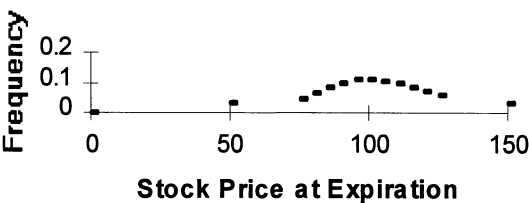
Let the probability density for prices at expiration be a histogram: in the interval,  $[s_j, s_{j+1}]$ ,  $S_T$  is uniformly distributed. Integrating a histogram gives a piecewise linear cdf, and integrating again gives a quadratic spline for call prices. To see this, write  $S_T$  as a mixture of uniform distributions,

$$F(s) = \sum_{j=1}^J p_j U_j(s)$$

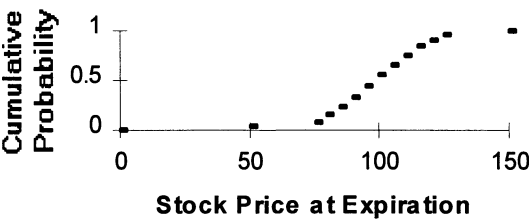
where,

$$U_j(s) = \begin{cases} 0 & s < s_j \\ \frac{s-s_j}{s_{j+1}-s_j} & s_j \leq s \leq s_{j+1} \\ 1 & s > s_j \end{cases}$$

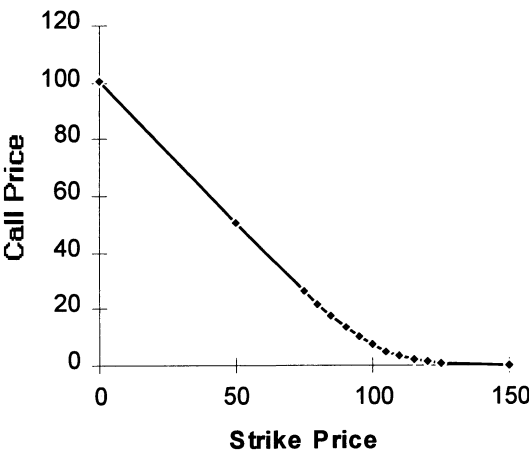
**Figure 1**  
**Discrete Probability**



**Discrete cdf**



**Call Prices with**  
**Discrete Probabilities**



Substituting into (1) gives the quadratic spline,

$$c(K) = c(s_j) + \sum_{j=1}^{j-1} p_j(K - s_j) + \frac{1}{2} p_j \left[ \frac{(K - s_j)^2}{(s_{j+1} - s_j)} \right] - (K - s_j), \quad s_j \leq K \leq s_{j+1}.$$

Conversely, if  $c(K)$  were a quadratic spline then the implied  $F$  would be a histogram density. This situation is illustrated in Figure 2.

### 2.3.3 Mixture models

The representation of the histogram as a mixture of uniform distributions can be extended directly to cases where the mixing distributions are not uniform. That is, now let

$$F(s) = \sum_{j=1}^J w_j F_j(s).$$

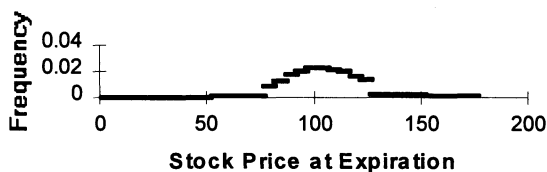
This says that  $S_T$  is, with probability  $w_j$ , the random variable  $S_j$  with cdf  $F_j$ . In the case where all the  $F_j$ s are lognormal with different variances, this corresponds to a lognormal stochastic volatility model in which Hull and White (1987) showed that call prices are the average of the call prices over the mixture. This extends to risk-neutral valuation with any mixing distributions, because

$$g \left( K : \sum_{j=1}^J w_j F_j(s) \right) = \sum_{j=1}^J w_j g(K : F_j(s)).$$

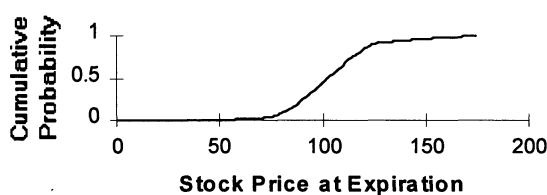
## 3 Estimating and bounding implied probability distributions

We first consider the case where call prices are arbitrage-free and hence convex, but where there are only a discrete number of strikes. Let  $c_i = c(K_i)$ ,  $i = 0, \dots, n$ , denote European call option prices on the same underlying asset with the same expiration date  $T$ , but different strike prices  $K_i$ , where  $K_i < K_{i+1}$  and  $K_0 = 0$ . The price of the zero price call is set equal to the current price of the stock,  $S_t$ . (If there are dividends then  $c(0) = e^{-\delta(T-t)} S_t$  where  $\delta$  is the payout rate through the expiration date). Assuming arbitrage-free valuation, the remaining call prices must satisfy,  $c_i = g(K_i : F)$ , for some unknown cdf  $F$ .

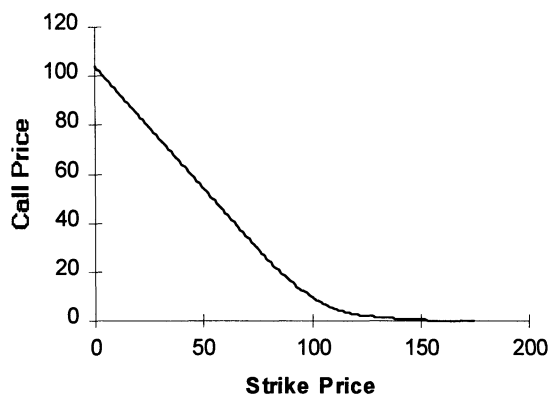
**Figure2**  
**Histogram Probability**



**Histogram cdf**



**Histogram Call Prices**



The estimation problem is to identify the  $F$  that generates the call prices  $(c_0, \dots, c_n)$ . Additional structure can be imposed by introducing restrictions on the set of allowed cdfs. When  $F$  is restricted to be lognormal with unknown variance the problem reduces to estimating the volatility smile. Less structure is imposed by Shimko (1993), who essentially estimates  $F$  and its associated density from a smoothed quadratic fit to the implied

volatilities. The method in Rubinstein (1994) is still more general as it essentially finds the risk-neutral density that is (least squares) closest to the lognormal density that could have generated call prices.

Consider the set of all cdf's that are consistent with the observed call prices, namely,

$$\{F \mid g(K_i : F) = c(K_i), i = 0, \dots, n\}.$$

If there was a continuum of strike prices, this set would consist of a single  $F$  and the implied cdf would be exactly

$$F(K) = e^{r(T-t)}[dc(K)/dK] + 1.$$

For the discrete strike case, let  $\hat{F}(K_i)$  be the analogous difference quotient,

$$\hat{F}(K_i) = e^{r(T-t)} \left[ \frac{c(K_{i+1}) - c(K_i)}{K_{i+1} - K_i} \right] + 1 \quad i = 0, \dots, n-1,$$

and set  $\hat{F}(K_n) = 1$ .

These  $\hat{F}(K_i)$  values can be used to bound the set of cdfs that could have generated call prices. Substitute  $g(K_i : F)$  for  $c(K_i)$  in the definition of  $\hat{F}(K_i)$  and use

$$F(K_i) \leq \frac{1}{(K_{i+1} - K_i)} \int_{K_i}^{K_{i+1}} F(s) ds \leq F(K_{i+1})$$

to show,

$$F(K_{i-1}) \leq \hat{F}(K_i) \leq F(K_i).$$

This yields upper and lower bounds on  $F$  at each strike price; namely,

$$\hat{F}(K_{i-1}) \leq F(K_i) \leq \hat{F}(K_i).$$

When the  $K_i$ s are close together this interpolation provides tight bounds for the allowed probabilities. When strike prices are far apart the bounds will be correspondingly large as it is then necessary to interpolate  $F$  over a large range of unobserved strike values.

Since  $F$  is nondecreasing, the  $\hat{F}(K_i)$  values can be used to bound the entire  $F$  function. Upper and lower cdfs are given by the discrete cdfs with jumps at the  $K_i$  values. Tighter bounds are given by assuming the underlying  $F$  is continuous. In this case the upper and lower cdfs are given by linearly interpolating between the  $\hat{F}(K_i)$  values.

Define an upper cdf by,

$$\hat{F}_U(K) = \hat{F}(K_i) + (K - K_i) \left[ \frac{\hat{F}(K_{i+1}) - \hat{F}(K_i)}{K_{i+1} - K_i} \right],$$

$$K_i \leq K \leq K_{i+1}, \quad i = 0, \dots, n-1,$$

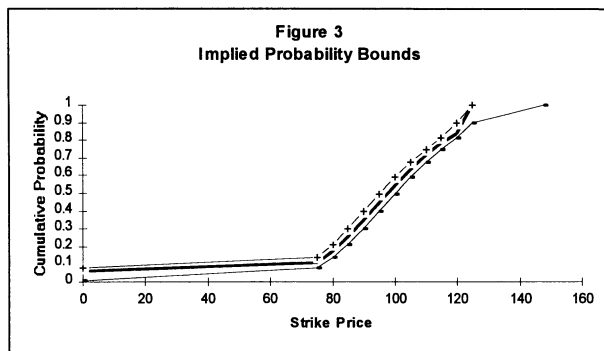
and a lower cdf by,

$$\hat{F}_L(K) = \hat{F}(K_{i-1}) + (K - K_{i-1}) \left[ \frac{\hat{F}(K_i) - \hat{F}(K_{i-1})}{K_i - K_{i-1}} \right],$$

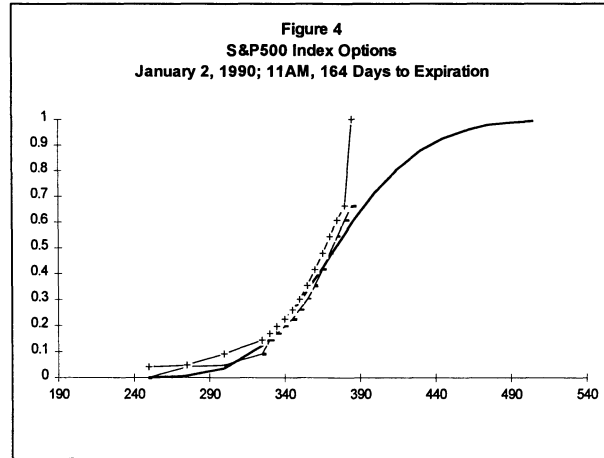
$$K_{i-1} \leq K \leq K_i, \quad i = 1, \dots, n.$$

These upper and lower cdfs bound the  $F$  that could have generated call prices, the only restriction being that the underlying cdf is continuous.

The bounds are illustrated in Figure 3. The data is from an example in Rubinstein (1994, p.781) in which call prices are generated by the Black Scholes model. For comparison the figure also shows a lognormal cdf. Since the call prices are generated from the Black Scholes model the lognormal cdf falls nicely between the bounding cdfs.



Another illustration, also from Rubinstein (1994, p. 784), is shown in Figure 4. Call prices are for the S&P500 Index at 11AM on January 2, 1990. The reference lognormal cdf is now seen to not fall within the bounding cdfs. The bounding cdfs are consistent with the estimated density function shown in Rubinstein in that the upper tail is much shorter and the left tail much longer than lognormal. Market prices imply a cdf with much greater chance of downward price movements than would be suggested by a lognormal cdf.



**Remark 1** *The distributions that fall between the upper and lower cdfs do not all have the same expectation. The allowed risk-neutral cdfs are those that fall within the bounds and which have their discounted expectation equal to the current stock price.*

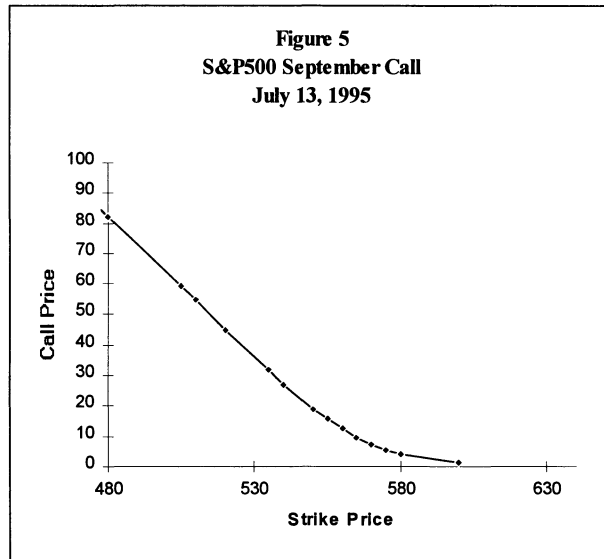
### 3.1 Estimation

Now let  $g(K_i : F)$  be a model for actual prices that are observed with error. The error term stands for all the non arbitrage reasons for differences between the call price and its  $g$  value. These reasons include: the bid-ask spread, nonsynchronous prices, and positive transaction costs.

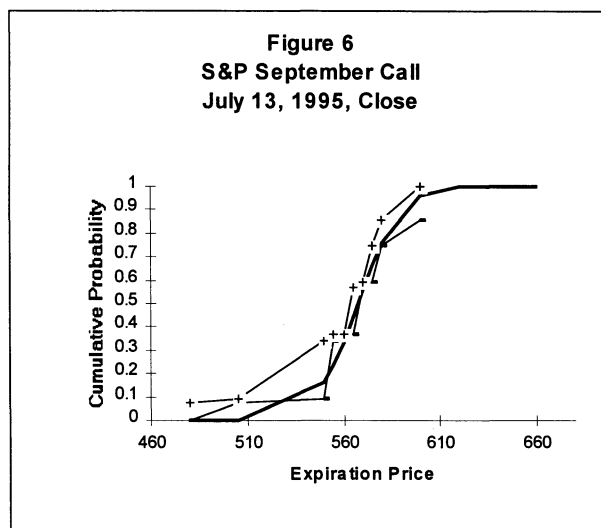
(It will be assumed that the difference between  $g$  and observed option prices does not depend on  $K$ . A more general analysis would permit the error variance to depend on trading volume, or, what is practically the same, the moneyness of the option,  $|S_t - K|$ .)

The presence of the error term means actual prices need not be convex. Hence, linear interpolation between adjacent strike prices need not yield a convex function, the resulting implied "cdf" based on the difference quotient need not be decreasing, and the implied density and probabilities could be negative.

Figure 5 shows call prices for the closing S&P500 call on July 13, 1995. Prices are almost convex; there is slight concavity in the deep in the money calls. (Note that, unlike the above example with 11AM prices, these closing prices are likely susceptible to slight departures from convexity on account of nonsynchronous prices near the close.) Since prices are nearly convex, the convex hull of observed prices is used for the values of  $g$  at the given strike prices.



The interest rate and dividend discount factors were estimated using put-call parity as in Shimko (1993). Bounds based on the convex hull of call prices are shown in Figure 6. The figure depicts bounding cdfs that are jumpier than shown in the previous figures, perhaps due to nonsynchronous closing prices. Similar to the other figures, however, there is a fat left-hand tail and a large difference from lognormality.





## Acknowledgements

I would like to thank Stan Pliska, Mark Rubinstein and a reviewer for helpful comments on an earlier version of the paper.

## References

- [1] Bassett, Gilbert and Roger Koenker (1982). An empirical quantile function for linear models with iid errors. *J. Am. Statist. Assoc.* **77**, 407-415.
- [2] Breeden, Douglas and Robert Litzenberger (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business* **51**, 621-651.
- [3] Cox, John, and Stephen Ross (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics* **3**, 145-166.
- [4] Cox, John and Mark Rubinstein (1985). *Options Markets*. New York: Prentice-Hall.
- [5] Fama, Eugene (1965). Mandelbrot and the stable paretian hypothesis. In *The Random Character of Stock Market Prices*, Ed. P. Cootner. MIT.
- [6] Figlewski, Stephen (1989). What does an option pricing model tell us about option prices? *Financial Analysts Journal*, September-October, 12-15.
- [7] Harrison, J. Michael and Stanley R. Pliska (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications* **11**, 215-260.
- [8] Hull, J.C. and A. White (1987). The pricing of options on assets with stochastic volatility. *Journal of Finance* **42**, 281-300.
- [9] Koenker, Roger and Gilbert W. Bassett Jr. (1978). Regression quantiles. *Econometrica* **46**, January, 33-50.
- [10] Mandelbrot, Benoit (1963). New methods in statistical economics. *Journal of Political Economy* **71**, 421-440.
- [11] Merton, Robert C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* **4**, 141-183.
- [12] Rockafellar, R. Tyrrell (1970). *Convex Analysis*. Princeton.
- [13] Rubinstein, Mark (1994). Implied binomial trees. *The Journal of Finance* **69**, pp.771-818.
- [14] Shimko, David (1993). Bounds of probability RISK **6**, 33-37.

# The marginal distributions of returns and volatility

Simon R. Hurst and Eckhard Platen

*The Australian National University, Canberra, Australia.*

*Abstract:* The aim of the paper is to identify a distribution that best fits empirical asset returns. By assuming a normal-variance mixture distribution for the returns, the distribution of volatility is implied.

*Key words:* Asset price model, leptokurtosis, quadratic variation process, mixture distributions, Student  $t$  distribution.

AMS subject classification: Primary 90A12; secondary 60J30, 62P20.

## 1 Introduction

The modelling of the stochastic process followed by the price of an asset is an important part of financial analysis. An understanding of this process is the first step to the pricing of derivative securities and general risk management. It is therefore important to identify a model for asset price processes which is consistent with their major empirical properties, such as heavy tailed return distributions, volatility clustering, long memory and persistence after volatility shocks. Previous approaches have typically concentrated on specific models, e.g. ARCH, and not succeeded so far to jointly model all of the major empirical properties. To attack this problem systematically we first study the marginal distributions of returns and volatility for market price indexes. Only after that we feel a substantial effort can be made to identify further evolutionary properties of volatility and asset price processes.

In this paper we compare various distributions to model the leptokurtic marginal distribution of asset returns. The distributions considered are: the normal (or Gaussian); the stable; the normal-lognormal mixture of Clark (1973); the generalised hyperbolic which include the Student  $t$ ,

the normal-inverse Gaussian mixture of Barndorff-Nielsen (1995), the hyperbolic of Eberlein and Keller (1995) and Küchler et al. (1995) and the variance-gamma (normal-gamma mixture) of Madan and Seneta (1990). These distributions are all mixtures of the normal distribution and differ only by their mixing volatility distributions.

It is crucial for any kind of serious risk analysis and management to emphasise the importance of correctly modelling the tail probabilities of returns. This is the reason why we will focus our comparative analysis on the identification of typical tail properties of index returns. Tests are performed on price indexes to directly determine a best marginal distribution for returns from the above mentioned alternatives. This distribution indirectly determines a best marginal distribution for the volatility. The best distribution for the index returns, with respect to the likelihood ratio test, turns out to be the Student  $t$  distribution. This distribution implies an inverted gamma distribution for the squared volatility. The Anderson and Darling (1952) test is also used to identify specifically the tail properties and additionally supports this result.

## 2 The class of generalised lognormal asset price models

Consider the *class of generalised lognormal models* for the asset price process  $\mathbf{S} = \{S(t), t \geq 0\}$  given by the Ito process with stochastic differential equation

$$dS(t) = \mu(t) S(t) dt + \sigma(t) S(t) dW(t), \quad (1)$$

for  $0 \leq t_0 \leq t < \infty$ . The stochastic process  $\mathbf{W} = \{W(t), t \geq 0\}$  represents the noise process which is assumed to be a *standard Wiener process* on a filtered probability space  $(\Omega, \mathcal{F}, \underline{\mathcal{F}} = \{\mathcal{F}_t\}_{t \geq 0}, P)$  fulfilling the usual conditions. We also have the *drift process*  $\mu = \{\mu(t), t \geq 0\}$  and the non-negative *volatility process*  $\sigma = \{\sigma(t), t \geq 0\}$ . These two processes may be constant, deterministic time-dependent or stochastic. In general we assume that they are  $\underline{\mathcal{F}}$ -adapted, right-continuous with left hand limits and that a unique, strong solution for (1) exists. The explicit solution of (1) for the asset price  $\mathbf{S}$  has the form

$$S(t) = S(t_0) \exp \left\{ \int_{t_0}^t (\mu(u) - \frac{1}{2} \sigma(u)^2) du + \int_{t_0}^t \sigma(u) dW(u) \right\}, \quad (2)$$

for  $0 \leq t_0 \leq t < \infty$ . Throughout this paper we will stay within the class of generalised lognormal asset price models.

Let us define the returns of the asset price process  $\mathbf{S}$ . We denote  $r_\Delta(t)$  to be the time  $t$  (continuously compounded) return of the asset price  $\mathbf{S}$  for

the interval  $[t, t + \Delta)$ . It is defined as

$$r_{\Delta}(t) = \log S(t + \Delta) - \log S(t). \quad (3)$$

When the drift  $\mu$  and the volatility  $\sigma$  are constant for all times  $t$ , the asset price model, defined by equation (1) or (2), is called the *classical lognormal model*. The Gaussian assumption of the theoretical return distribution of this classical model seriously restricts its shape, especially the tail thickness. Asset returns are usually observed to have leptokurtic empirical distributions. That is, they have heavier tails and have a more pronounced peak around the mode than a normal distribution.

We denote the *log asset price process* by  $\mathbf{L} = \{L(t) = \log S(t), t \geq 0\}$ . The *quadratic variation process*  $\langle \mathbf{L} \rangle = \{\langle L \rangle(t), t \geq 0\}$  is given by (see e.g. Jacod and Shiryaev, 1987, §4e)

$$\langle L \rangle(t) = \int_0^t \sigma(u)^2 du, \quad (4)$$

for a generalised lognormal model.

We define the *empirical quadratic variation process*  $\langle \mathbf{L} \rangle_{\Delta} = \{\langle L \rangle_{\Delta}(t), t \geq 0\}$ , based on time steps of length  $\Delta$ , of the log asset price process  $\mathbf{L}$  to be

$$\langle L \rangle_{\Delta}(t) = \sum_{j=0}^{n-1} r_{\Delta}(j \Delta)^2, \quad (5)$$

where  $n \Delta \leq t < (n + 1) \Delta$ , for some  $n \in \mathbf{N}$ , and  $r_{\Delta}(\cdot)$  are the returns defined in (3). Note that the empirical quadratic variation process  $\langle \mathbf{L} \rangle_{\Delta}$  is an estimate for the true underlying quadratic variation process  $\langle \mathbf{L} \rangle$ . It converges (under rather general assumptions)  $P$ -a.s. to  $\langle \mathbf{L} \rangle$  as  $\Delta \rightarrow 0$ .

The daily empirical quadratic variation processes of the log indexes are shown in Figure 1. This figure indicates that the quadratic variation processes  $\langle \mathbf{L} \rangle$  are stochastic non-decreasing processes. Consequently it follows from (4) that the volatility  $\sigma$  is a stochastic process. One can say it is the stochastic nature of volatility which makes the distribution of returns leptokurtic. Below it will be generally shown that the theoretical return distribution in a generalised lognormal model is leptokurtic when the volatility  $\sigma$  is stochastic.

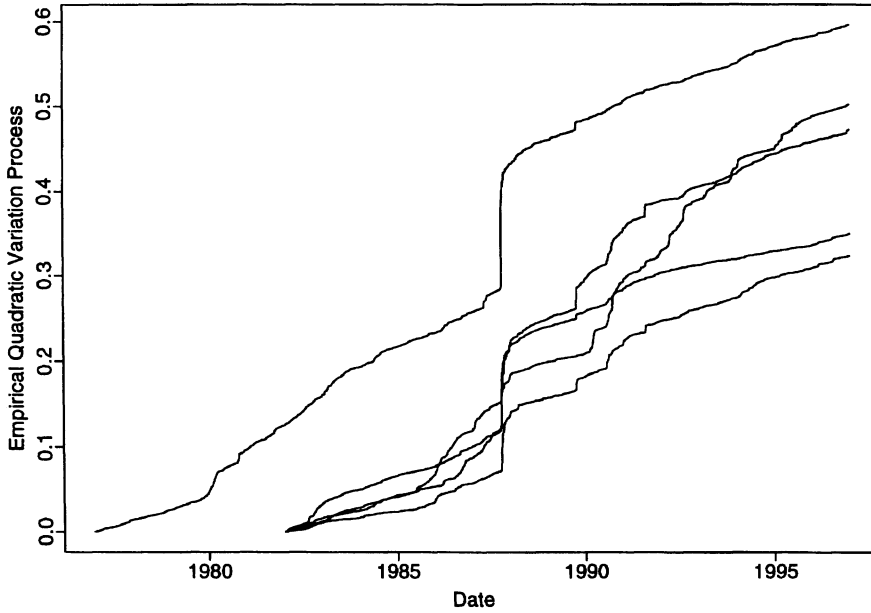


Figure 1: Daily empirical quadratic variation processes for the log market price indexes.

### 3 Mixing distributions

Consider the time discretisation

$$0 = t_0 \leq t_1 \leq t_2 \leq \dots, \quad (6)$$

where  $t_i = i\Delta$  for all  $i \in \mathbf{N}$  and  $\Delta > 0$ . Define  $i_t$  as the largest integer  $i$  such that  $t_i$  is less than or equal to  $t$ , i.e.  $i_t = \max\{i \in \mathbf{N} : t_i \leq t\}$ . We obtain the important class of discrete time generalised lognormal asset price models by keeping the drift  $\mu$  and the volatility  $\sigma$  piece-wise constant over the discretisation intervals. A discrete time generalised lognormal model's log asset price process is then given by  $\mathbf{L}_\Delta = \{L_\Delta(t) = L_\Delta(t_{i_t}), t \geq 0\}$ , where

$$L_\Delta(t_{i+1}) = L_\Delta(t_i) + (\mu(t_i) - \frac{1}{2}\sigma(t_i)^2)\Delta + \sigma(t_i)\Delta W(t_i), \quad (7)$$

for  $i \in \mathbf{N}$  and where  $\Delta W(t_i)$ ,  $i = 0, 1, 2, \dots$ , are independent and identically distributed normal random variables with zero mean and variance  $\Delta$ . The stochastic difference equation can be interpreted as that of an Euler approximation for a certain log asset price process  $\mathbf{L}$  as defined earlier.

Note that the discrete time log asset price process  $\mathbf{L}_\Delta$  then converges to its continuous time limit  $\mathbf{L}$  as the time step size  $\Delta$  tends to zero (see e.g. Kloeden and Platen, 1992).

Consider now the random variable  $X$ . We denote its distribution function (d.f.) by  $F_X(\cdot)$ , its characteristic function (c.f.) by  $\phi_X(\cdot)$ , and its probability density function (p.d.f.) by  $f_X(\cdot)$  if it exists. We also denote the  $n$ th moment of  $X$  by  $m_{X,n} = E(X^n)$ , where  $E(\cdot)$  denotes the expectation operator. It is well known that the mean  $\mu_X = E(X)$ , variance  $\sigma_X^2 = E((X - \mu_X)^2)$ , skewness  $\beta_X = E((X - \mu_X)^3)/\sigma_X^3$  and kurtosis  $\kappa_X = E((X - \mu_X)^4)/\sigma_X^4$  are respectively measures of the location, variability, degree of asymmetry and tail thickness/peakness of the distribution of  $X$ . Note that the moments  $m_{X,n}$ ,  $n \in \mathbf{N}$ , can be calculated by using the c.f.  $\phi_X(\cdot)$  with the well known formula

$$m_{X,n} = (-i)^n \phi_X^{(n)}(0), \quad (8)$$

where  $\phi_X^{(n)}(\cdot)$  denotes the  $n$ th derivative of the c.f. The mean, variance, skewness and kurtosis can be calculated from the moments.

The volatility  $\sigma$  is an unobservable quantity. As such, the quantification of the distribution of volatility can not be directly obtained. However we can obtain it indirectly via the distribution of returns. Equations (3) and (7) give the returns for the discrete time generalised lognormal model as

$$r_\Delta(t_i) = (\mu(t_i) - \frac{1}{2} \sigma(t_i)^2) \Delta + \sigma(t_i) \Delta W(t_i), \quad (9)$$

for  $i \in \mathbf{N}$ . The return  $r_\Delta(t_i)$ , conditioned on the random variable  $\sigma(t_i)^2$ , is a normal distributed random variable with mean  $(\mu(t_i) - \frac{1}{2} \sigma(t_i)^2) \Delta$  and variance  $\sigma(t_i)^2 \Delta$ . The drift coefficient  $\mu(t_i)$  may depend on the random variable  $\sigma(t_i)^2$  so we write the conditional mean as  $\xi(t_i, \sigma(t_i)^2) = \mu(t_i) - \frac{1}{2} \sigma(t_i)^2$  to denote this possible dependence.

Let us assume some properties for the conditional mean  $\xi(t_i, \sigma(t_i)^2) \Delta$  and the conditional variance  $\sigma(t_i)^2 \Delta$  of the return  $r_\Delta(t_i)$  given in (9):

First we assume that the volatility  $\sigma$  is a stationary process (see e.g. Feller, 1966, §III.7). Then  $\sigma(t)$  is identically distributed according to the invariant distribution of  $\sigma$  for any given time  $t$ . The stationarity is a reasonable assumption because if we look at the quadratic variation processes in Figure 1 we notice that the processes are similar over the entire time period. They could be described as having linear trends interspersed with strongly increasing periods (very volatile) which slowly revert back towards another linear trend with the same slope. This indicates that the volatility is a stationary process.

We also make the simplifying assumption that the mean has the structure

$$\xi(t_i, \sigma(t_i)^2) \Delta = (\eta + \varrho \sigma(t_i)^2) \Delta, \quad (10)$$

for some constants  $\eta$  and  $\varrho$ .

These two assumptions imply the following relations for the unconditional distribution of returns  $r_\Delta(\cdot)$ . It follows that their d.f. is

$$F_{r_\Delta}(x) = \int_0^\infty \Phi\left(\frac{x - (\eta + \varrho u) \Delta}{\sqrt{u} \Delta}\right) dF_{\sigma^2}(u) \quad (x \in \mathbf{R}), \quad (11)$$

where  $\Phi(\cdot)$  is the standard normal d.f. The corresponding p.d.f. is

$$f_{r_\Delta}(x) = \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{1}{\sqrt{u} \Delta} \exp\left\{-\frac{(x - (\eta + \varrho u) \Delta)^2}{2 u \Delta}\right\} dF_{\sigma^2}(u) \quad (x \in \mathbf{R}), \quad (12)$$

if it exists, and the corresponding c.f. is

$$\phi_{r_\Delta}(\theta) = \int_0^\infty \exp\left\{i(\eta + \varrho u) \Delta \theta - \frac{1}{2} u \Delta \theta^2\right\} dF_{\sigma^2}(u) \quad (\theta \in \mathbf{R}). \quad (13)$$

The unconditional distribution is called a *mixture distribution* and the distribution of  $\sigma^2$  is called the *mixing distribution* (see e.g. Feller, 1966, §II.5).

Feller (1966), §XVII.3(i), gives the general representation of the c.f.  $\phi_{\sigma^2}(\cdot)$  for the non-negative random variable  $\sigma^2$ . Then the moments  $m_{\sigma^2, n}$ ,  $n \in \mathbf{N}$ , can be calculated by using (8) and this representation of the c.f.  $\phi_{\sigma^2}(\cdot)$ . The mean  $\mu_{\sigma^2}$ , variance  $\sigma_{\sigma^2}^2$ , skewness  $\beta_{\sigma^2}$  and kurtosis  $\kappa_{\sigma^2}$  can consequently be calculated. It is easily shown from these values that  $\mu_{\sigma^2}, \sigma_{\sigma^2}^2, \beta_{\sigma^2} > 0$  and  $\kappa_{\sigma^2} > 3$ , if  $\sigma^2$  is not deterministic. That is, the random variable  $\sigma^2$  is positively skewed and leptokurtic.

Let us now compute the mean  $\mu_{r_\Delta}$ , variance  $\sigma_{r_\Delta}^2$ , skewness  $\beta_{r_\Delta}$  and kurtosis  $\kappa_{r_\Delta}$  for the return  $r_\Delta(\cdot)$  as important measures required to understand the distributional properties of asset prices. Formulae for these measures are obtained via equations (8) and (13) giving

$$\begin{aligned} \mu_{r_\Delta} &= \eta \Delta + \varrho \mu_{\sigma^2} \Delta, \\ \sigma_{r_\Delta}^2 &= \mu_{\sigma^2} \Delta + \varrho^2 \sigma_{\sigma^2}^2 \Delta^2, \\ \beta_{r_\Delta} &= \frac{3 \varrho \sigma_{\sigma^2}^2 \Delta^2 + \varrho^3 \beta_{\sigma^2} \sigma_{\sigma^2}^3 \Delta^3}{\sigma_{r_\Delta}^3}, \\ \kappa_{r_\Delta} &= \frac{3 (\mu_{\sigma^2}^2 + \sigma_{\sigma^2}^2) \Delta^2 + 6 \varrho^2 (\mu_{\sigma^2} \sigma_{\sigma^2}^2 + \beta_{\sigma^2} \sigma_{\sigma^2}^3) \Delta^3 + \varrho^4 \kappa_{\sigma^2} \sigma_{\sigma^2}^4 \Delta^4}{\sigma_{r_\Delta}^4} \end{aligned} \quad (14)$$

It now easily follows that if  $\sigma^2$  is stochastic, i.e.  $\sigma_{\sigma^2}^2 > 0$ , then  $\kappa_{r_\Delta} > 3$ . We also get that  $\text{sign}(\beta_{r_\Delta}) = \text{sign}(\varrho)$ , by using (14) and the above fact that  $\sigma^2$

is positively skewed and leptokurtic, i.e.  $\beta_{\sigma^2} > 0$ ,  $\kappa_{\sigma^2} > 3$ . The returns  $r_{\Delta}(\cdot)$  are therefore leptokurtic and skewed in the direction of the sign of  $\varrho$ . The above relations show that the stochastic nature of volatility implies a leptokurtic distribution for the returns.

We state that in our analysis of price indexes in Section 5 the empirical return distributions are fairly symmetrical. Consequently to simplify the analysis, we assume that the distribution of asset returns is *symmetric*. This means we explicitly assume that  $\varrho = 0$  from this point onward.

## 4 Marginal distributions of asset returns

As already mentioned we focus our analysis on the marginal distribution of asset returns. The marginal distributions for the returns in the models which we examine in this section are mixtures of the normal distribution. They differ by their mixing distribution  $\sigma^2$ . As discussed above, this implies a leptokurtic distribution for the returns if  $\sigma^2$  is random. The models can be characterised by *either* its mixing distribution for  $\sigma^2$  *or* equivalently by its mixture distribution for  $r_{\Delta}$ , since each are related by equations (11), (12) and (13).

Below we briefly characterise the different models by only giving the probability density function  $f_{r_{\Delta}}(\cdot)$  or the characteristic function  $\phi_{r_{\Delta}}(\cdot)$  for the marginal distribution of the returns  $r_{\Delta}(\cdot)$ . A more detailed treatment can be found in Hurst, Platen and Rachev (1996).

It must be emphasised that we do not intend to similarly model the asset price process as having independent increments as the following models do. The i.i.d. modelling assumption is inconsistent with the well known properties of volatility clustering, long-memory and persistence for the asset prices. We will interpret the result more correctly as an identification of the marginal distribution for asset returns.

### 4.1 The Mandelbrot and Fama logstable model

Mandelbrot (1963, 1967) and Fama (1963, 1965) proposed returns to be distributed with an  $\alpha$ -stable distribution. This occurs when the stationary distribution of  $\sigma^2$  is a maximally skewed  $\alpha/2$ -stable distribution with  $\alpha \in (0, 2)$  (see e.g. Mandelbrot and Taylor, 1967). The c.f. of the returns  $r_{\Delta}(\cdot)$  is

$$\phi_{r_{\Delta}}(\theta) = \exp \{i \eta \Delta \theta - c^{\alpha} \Delta |\theta|^{\alpha}\} \quad (\theta \in \mathbf{R}). \quad (15)$$

The parameter  $\alpha$  is called the *index of stability* and is a shape parameter for the distribution, the smaller  $\alpha$ , the larger the tail thickness. This model also implies infinite variance and infinite kurtosis for the return distribution.



## 4.2 The Clark model

Clark (1973) proposed the change in the *asset price* process to be distributed with a normal-lognormal mixture distribution. Prices have to be positive and represent growth processes therefore we are interested in the change in the *log asset price* process, i.e. returns. We modify Clark's model here and propose the returns to be distributed with a normal-lognormal mixture distribution. In this modified version of the Clark model the stationary distribution of  $\sigma^2$  is a lognormal distribution. The p.d.f. of the returns  $r_\Delta(\cdot)$  is

$$f_{r_\Delta}(x) = \frac{1}{2\pi\varphi\sqrt{\Delta}} \int_0^\infty u^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\left(\frac{(x-\eta\Delta)^2}{u\Delta} + \frac{(\log u - \log c^2 + \frac{1}{2}\varphi^2)^2}{\varphi^2}\right)\right\} du \quad (x \in \mathbf{R}), \quad (16)$$

The parameter  $\varphi$  is a shape parameter for the distribution. The return distribution for this model has kurtosis  $\kappa_{r_\Delta} = 3\exp(\varphi^2)$ .

## 4.3 The log symmetric generalised hyperbolic model

Various authors (e.g. Praetz, 1972; Blattberg and Gonedes, 1974; Madan and Seneta, 1990; Barndorff-Nielsen 1995; Eberlein and Keller 1995; K  chler et al., 1995) have proposed returns to be distributed within the class of the generalised hyperbolic distributions. We consider the more restrictive *symmetric* generalised hyperbolic distributions for the return distribution. These distributions result when the stationary distribution of  $\sigma^2$  is a generalised inverse Gaussian distribution. The p.d.f. of the returns  $r_\Delta(\cdot)$  is

$$f_{r_\Delta}(x) = \frac{1}{\delta\sqrt{\Delta}K_\lambda(\alpha\delta)} \sqrt{\frac{\alpha\delta}{2\pi}} \left(1 + \frac{(x-\eta\Delta)^2}{\delta^2\Delta}\right)^{\frac{1}{2}(\lambda-\frac{1}{2})} K_{\lambda-\frac{1}{2}}\left(\alpha\delta\sqrt{1 + \frac{(x-\eta\Delta)^2}{\delta^2\Delta}}\right) \quad (x \in \mathbf{R}). \quad (17)$$

where  $K_\lambda(\cdot)$  is the modified Bessel function of the third kind with index  $\lambda$ ,  $\lambda \in \mathbf{R}$  and  $\alpha, \delta \geq 0$ . In addition  $\alpha \neq 0$  if  $\lambda \geq 0$  and  $\delta \neq 0$  if  $\lambda \leq 0$ . The parameters  $\lambda$  and  $\bar{\alpha} = \alpha\delta$  are invariant shape parameters. The return distribution for this model has kurtosis  $\kappa_{r_\Delta} = 3K_\lambda(\bar{\alpha})K_{\lambda+2}(\bar{\alpha})/K_{\lambda+1}(\bar{\alpha})^2$ .

In the following we briefly consider special parameterisations of this model which have previously been considered by other authors.

#### 4.4 The log Student $t$ model

Praetz (1972) and Blattberg and Gonedes (1974) proposed that the returns should be distributed with a Student  $t$  distribution with degrees of freedom  $\nu > 0$ . This occurs when the shape parameters  $\lambda = -\frac{1}{2}\nu < 0$  and  $\bar{\alpha} = 0$ , i.e.  $\alpha = 0$ , and the parameter  $\delta = c\sqrt{\nu}$ . The p.d.f. of the returns  $r_\Delta(\cdot)$  is

$$f_{r_\Delta}(x) = \frac{\Gamma(\frac{1}{2}\nu + \frac{1}{2})}{c\sqrt{\pi\nu\Delta}\Gamma(\frac{1}{2}\nu)} \left(1 + \frac{(x - \eta\Delta)^2}{c^2\nu\Delta}\right)^{-\frac{1}{2}\nu - \frac{1}{2}} \quad (x \in \mathbf{R}). \quad (18)$$

The degrees of freedom  $\nu$  is the shape parameter for the distribution. The return distribution for this model has kurtosis  $\kappa_{r_\Delta} = 3(\nu - 2)/(\nu - 4)$ , for  $\nu > 4$ , and is infinite otherwise.

#### 4.5 The Barndorff-Nielsen log normal\inverse Gaussian model

Barndorff-Nielsen (1995) proposed returns to be distributed with a normal-inverse Gaussian mixture distribution. This occurs when the shape parameter  $\lambda = -\frac{1}{2}$ . The p.d.f. of the returns  $r_\Delta(\cdot)$  is

$$f_{r_\Delta}(x) = \frac{\sqrt{\bar{\alpha}} \exp\{\bar{\alpha}\}}{c\sqrt{\Delta}\pi} \left(1 + \frac{(x - \eta\Delta)^2}{\bar{\alpha}c^2\Delta}\right)^{-\frac{1}{2}} K_1\left(\bar{\alpha}\sqrt{1 + \frac{(x - \eta\Delta)^2}{\bar{\alpha}c^2\Delta}}\right) \quad (x \in \mathbf{R}). \quad (19)$$

The parameter  $\bar{\alpha}$  is the shape parameter for the distribution. The return distribution for this model has kurtosis  $\kappa_{r_\Delta} = 3(1 + 1/\bar{\alpha})$ .

#### 4.6 The log hyperbolic model

Eberlein and Keller (1995) and K  chler et al. (1995) proposed returns to be distributed with a hyperbolic distribution. This occurs when the shape parameter  $\lambda = 1$ . The p.d.f. of the returns  $r_\Delta(\cdot)$  is

$$f_{r_\Delta}(x) = \frac{1}{2\delta\sqrt{\Delta}K_1(\bar{\alpha})} \exp\left\{-\bar{\alpha}\sqrt{1 + \frac{(x - \eta)^2}{\delta^2\Delta}}\right\} \quad (x \in \mathbf{R}). \quad (20)$$

The parameter  $\bar{\alpha}$  is the shape parameter for the distribution. The return distribution for this model has kurtosis  $\kappa_{r_\Delta} = 3K_1(\bar{\alpha})K_3(\bar{\alpha})/K_2(\bar{\alpha})^2$ .

#### 4.7 The Madan and Seneta log variance gamma model

Madan and Seneta (1990) proposed returns to be distributed with a normal-gamma mixture distribution. This occurs when the shape parameters  $\lambda > 0$  and  $\bar{\alpha} = 0$ , i.e.  $\delta = 0$ . The p.d.f. of the returns  $r_\Delta(\cdot)$  is

$$f_{r_\Delta}(x) = \frac{\sqrt{\lambda}}{c\sqrt{\Delta}\pi\Gamma(\lambda)2^{\lambda-1}} \cdot \left(\sqrt{2\lambda} \frac{|x - \eta\Delta|}{c\sqrt{\Delta}}\right)^{\lambda-\frac{1}{2}} K_{\lambda-\frac{1}{2}}\left(\sqrt{2\lambda} \frac{|x - \eta\Delta|}{c\sqrt{\Delta}}\right) \quad (x \in \mathbf{R}). \quad (21)$$

The parameter  $\lambda$  is the shape parameter for the distribution. The return distribution for this model has kurtosis  $\kappa_{r_\Delta} = 3(1 + 1/\lambda)$ .

### 5 Analysis of major world market indexes

The empirical analysis will be performed on market indexes from the United States of America, Japan, Germany, Switzerland and Australia. The Australian index is calculated by Datastream International and the other indexes are calculated by Morgan Stanley Capital International. The data for these indexes are daily data for the 15 years from the beginning of 1982 to the end of 1996, except for 20 years of Australian index data which start from the beginning of 1977. We note that all of these indexes include the stock market crash of October 1987.

In Table 1 we display the results of our analysis. Under the heading of *Empirical Model* we show the total number of daily returns  $n$  and the sample measure of kurtosis  $\hat{\kappa}_{r_\Delta}$ . Also included in this table is the sample measure of kurtosis  $\hat{\kappa}_{r_\Delta}^*$  corresponding to the data with the largest absolute return removed.

The two sample measures of kurtosis,  $\kappa_{r_\Delta}$  and  $\hat{\kappa}_{r_\Delta}^*$ , indicate that the index returns are highly leptokurtic and hence are very heavy tailed. In fact they are so large that higher moments (including the fourth moment) may be unbounded for the index returns, i.e. be infinite. In this case the sample measure of kurtosis would be unstable. This is what we observe by removing one extreme observation from the sample, the sample kurtosis changes significantly for each index. We also observe this property by a plot of the sample measure of kurtosis against the sample size which is not shown here. It is therefore important in our analysis to concentrate on the *entire distribution* and not just on a *single statistic* (in particular the possibly unbounded sample kurtosis) to identify a good or best model, or more precisely a best marginal distribution. We will therefore base our final judgement on the likelihood ratio test.

It can be shown that all of the models in Section 4 include the classical lognormal model as a specific or limiting case. Consequently we can test if each model is significantly better than the classical lognormal model by using the likelihood ratio test (see e.g. Rao, 1973, §6e.2). Define the likelihood ratio

$$\Lambda = \frac{\mathcal{L}_{\lognormal}}{\mathcal{L}_{other}}, \quad (22)$$

where  $\mathcal{L}_{\lognormal}$  is the likelihood value of the classical lognormal model and  $\mathcal{L}_{other}$  is the likelihood value of the other model we are testing. The asymptotic distribution of  $-2\log\Lambda$  is chisquare with degrees of freedom equal to the difference in the number of parameters between the two models. Large values of  $-2\log\Lambda$  indicate that the model under consideration is significantly better (explains more) than the classical lognormal model. We choose the model with the significantly <sup>1</sup> largest value of  $-2\log\Lambda$  to be the *best* model. Intuitively, this is the model which has the largest probability for the returns and therefore is adding the most information to the classical lognormal model with the minimum number of parameters.

Another way of comparing the models, especially the tail properties, is to statistically determine how close the empirical distribution function and the model's distribution function are for the returns. We use the Anderson and Darling (1952) test here. This test increases the power of the more commonly used Kolmogorov test in the tails of the distribution by using the properly *weighted* test statistic given by

$$AD = \max_{x \in \mathbf{R}} \frac{|F_e(x) - F_m(x)|}{\sqrt{F_m(x)(1 - F_m(x))}}, \quad (23)$$

where  $F_e(\cdot)$  is the empirical distribution function and  $F_m(\cdot)$  is the model's distribution function. A good (bad) fit is indicated by a small (large) difference between the two distribution functions and hence a small (large) value of the test statistic  $AD$ .

The parameters for each model of Section 4 are estimated by the maximum likelihood method. For each model we display in Table 1 the estimated shape parameter(s), the corresponding kurtosis  $\kappa_{r_\Delta}$ , the likelihood ratio test value  $-2\log\Lambda$  and the Anderson-Darling test statistic  $AD$ .

For all of the models the likelihood ratio test values  $-2\log\Lambda$  are extremely large indicating that all of the models from Section 4 are significantly better than the classical lognormal model. In Table 1 we have highlighted the most significant likelihood ratio value  $-2\log\Lambda$  for each in-

---

<sup>1</sup>The log symmetric generalised hyperbolic model has an extra parameter than all of the other models and so it has to be accounted for correctly.

dex. It is clear that the log Student  $t$  model is the best model for all five indexes.

Table 1: Results for the quantification of the marginal distributions of returns and volatility.

| Model                                  | Statistic                  | Country         |                 |                 |                  |                  |
|----------------------------------------|----------------------------|-----------------|-----------------|-----------------|------------------|------------------|
|                                        |                            | Australia       | Germany         | Japan           | Switzerland      | USA              |
| Empirical                              | $n$                        | 5052            | 3761            | 3723            | 3761             | 3803             |
|                                        | $\hat{\kappa}_{r\Delta}$   | 125.7994        | 16.7455         | 21.4201         | 26.037           | 93.5112          |
|                                        | $\hat{\kappa}_{r\Delta}^*$ | 9.4216          | 11.1454         | 11.1512         | 17.5189          | 11.3116          |
| Normal                                 | $\kappa_{r\Delta}$         | 3               | 3               | 3               | 3                | 3                |
|                                        | $AD$                       | 2.67e+81        | 1.47e+14        | 4.34e+19        | 3.88e+18         | 1.45e+59         |
| Stable                                 | $\alpha$                   | 1.8055          | 1.7237          | 1.6099          | 1.6931           | 1.6878           |
|                                        | $\kappa_{r\Delta}$         | $\infty$        | $\infty$        | $\infty$        | $\infty$         | $\infty$         |
|                                        | $-2 \log \Lambda$          | 1433.665        | 828.7447        | 1008.4832       | 1132.1436        | 1180.4462        |
|                                        | $AD$                       | <b>0.045991</b> | 0.061149        | 0.057997        | <b>0.060315</b>  | <b>0.056682</b>  |
| Clark                                  | $\varphi$                  | 0.8452          | 0.9182          | 1.0643          | 0.9833           | 1.0004           |
|                                        | $\kappa_{r\Delta}$         | 6.1291          | 6.9712          | 9.3127          | 7.889            | 8.1616           |
|                                        | $-2 \log \Lambda$          | 1416.0134       | 806.4511        | 1034.6007       | 1101.4388        | 1229.0101        |
|                                        | $AD$                       | 43.745          | 0.29045         | 0.22173         | 0.35477          | 3.5355           |
| Symmetric<br>Generalised<br>Hyperbolic | $\lambda$                  | -2.2721         | -1.9253         | -1.4109         | -1.753           | -1.7799          |
|                                        | $\bar{\alpha}$             | 2.6803e-06      | 9.3886e-07      | 0.20115         | 1.7939e-06       | 5.8455e-07       |
|                                        | $\kappa_{r\Delta}$         | 14.0143         | 263.349         | 19.1107         | 6581.6199        | 6135.9287        |
|                                        | $-2 \log \Lambda$          | 1449.143        | 836.0898        | 1048.1408       | 1138.6009        | 1231.9909        |
|                                        | $AD$                       | 0.37820         | 0.057845        | 0.054036        | 0.081052         | 0.12035          |
| Student $t$                            | $\nu$                      | 4.5441          | 3.8506          | 3.0687          | 3.5061           | 3.5598           |
|                                        | $\kappa_{r\Delta}$         | 14.0268         | $\infty$        | $\infty$        | $\infty$         | $\infty$         |
|                                        | $-2 \log \Lambda$          | <b>1449.143</b> | <b>836.0898</b> | <b>1047.256</b> | <b>1138.6009</b> | <b>1231.9909</b> |
|                                        | $AD$                       | 0.37838         | <b>0.057845</b> | <b>0.030761</b> | 0.081066         | 0.12073          |
| Normal<br>Inverse<br>Gaussian          | $\bar{\alpha}$             | 0.97355         | 0.80127         | 0.52894         | 0.6519           | 0.6359           |
|                                        | $\kappa_{r\Delta}$         | 6.0815          | 6.7441          | 8.6717          | 7.602            | 7.7178           |
|                                        | $-2 \log \Lambda$          | 1388.7137       | 800.6066        | 1033.4658       | 1092.0354        | 1209.8897        |
|                                        | $AD$                       | 7808.0          | 0.71372         | 0.62053         | 1.0496           | 105.02           |
| Hyperbolic                             | $\bar{\alpha}$             | 0.72732         | 0.57797         | 0.25477         | 0.42248          | 0.28722          |
|                                        | $\kappa_{r\Delta}$         | 5.1335          | 5.3104          | 5.7425          | 5.5127           | 5.6981           |
|                                        | $-2 \log \Lambda$          | 1343.293        | 755.1277        | 970.3204        | 1023.5112        | 1173.6493        |
|                                        | $AD$                       | 6.54e+06        | 8.5788          | 26.189          | 31.885           | 88645            |
| Variance<br>Gamma                      | $\lambda$                  | 1.481           | 1.3375          | 1.0212          | 1.1912           | 1.1225           |
|                                        | $\kappa_{r\Delta}$         | 5.0256          | 5.243           | 5.9376          | 5.5185           | 5.6726           |
|                                        | $-2 \log \Lambda$          | 1322.8124       | 739.5928        | 965.7311        | 1007.2743        | 1165.7439        |
|                                        | $AD$                       | 3.15e+07        | 11.613          | 19.632          | 36.099           | 1.00e+05         |

Mixed results are obtained when we use the Anderson and Darling test. Smaller  $AD$  values indicate a better fit. There is a mixture between the log Student  $t$  model and the log stable model as to which model is the best. We note that the October 1987 crash return is having a great influence on the test statistic and therefore seems to bias towards heavy tailed distributions. We consider this test only as an additional check for identifying the appropriate tail properties. From our point of view the likelihood ratio test provides the most objective basis for a comparison between the marginal return distributions.

Based on the above results we conclude that the three parameter Student  $t$  distribution is the best marginal distribution for index returns. It is closely followed by the four parameter symmetric generalised hyperbolic distribution, which for all but the Japanese index turns out to be exactly the Student  $t$  distribution. The stable, Clark and normal\\inverse Gaussian distributions can be described as distributions which do not explain the tail properties accurately enough. The stable distribution overestimates the tail thickness whereas the Clark and normal\\inverse Gaussian distributions both underestimate it. The hyperbolic and variance gamma distributions are poor and dramatically underestimate the tail properties.

Some readers may argue that for each index the large negative return caused by the stock market crash of October 1987 is an outlier and is therefore influencing our results. That is, in favour of a model with very heavy tails opposed to one with less heavy tails. We would like to point out that from the paper on extreme value theory for asset price returns by Longin (1996) the stock market crash of October 1987 is dismissed as an outlier, i.e. it is consistent with the rest of the data. For principle reasons we do not like to exclude the extreme events such as stock market crashes from our samples because it is this feature (namely tail heaviness) we are explicitly emphasising to model in a consistent way. However to provide a view on the robustness of our results we removed the large negative return caused by the stock market crash of October 1987 and repeated our study. As to be expected by removing extreme events, the distributions with thinner tails improve whereas the stable distribution, which has the heaviest tails, gets worse (the results can be obtained from the authors). The Student  $t$  distribution is still the best distribution from the alternatives considered.

## 6 Conclusion

The Student  $t$  distribution has been shown above to be the best marginal distribution for index returns, with respect to the likelihood ratio test. It implies an inverted gamma distribution for the marginal distribution of the squared volatility  $\sigma^2$ . This distributional property can now be exploited to identify possible dynamics of the volatility process  $\sigma$  and hence the evolution of the asset price process.

## References

- [1] Anderson, T. W. and D. A. Darling (1952). Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *Ann. Math.*

- Statist.* **23**, 193–212.
- [2] Barndorff-Nielsen, O. E. (1995). Normal Inverse Gaussian processes and the modelling of stock returns. Research Report 300, University of Aarhus.
  - [3] Blattberg, R. C. and N. Gonedes (1974). A comparison of the stable and Student distributions as statistical models for stock prices. *Journal of Business* **47**, 244–280.
  - [4] Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* **41**, 135–159.
  - [5] Eberlein, E. and U. Keller (1995). Hyperbolic distributions in finance. *Bernoulli* **1**, 281–299.
  - [6] Fama, E. F. (1963). Mandelbrot and the stable paretian hypothesis. *Journal of Business* **36**, 420–429.
  - [7] Fama, E. F. (1965). The behavior of stock-market prices. *Journal of Business* **38**, 34–105.
  - [8] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Volume 2. New York: Wiley.
  - [9] Hurst, S. R., E. Platen, and S. T. Rachev (1996). Subordinated market index models: A comparison. Research Report FMRR 002–96, The Australian National University. *Financial Engineering and the Japanese Markets*. To appear.
  - [10] Jacod, J. and A. N. Shiryaev (1987). *Limit Theorems for Stochastic Processes*. Berlin: Springer-Verlag.
  - [11] Kloeden, P. E. and E. Platen (1992). *Numerical Solution of Stochastic Differential Equations*. Berlin: Springer-Verlag.
  - [12] Küchler, U., K. Neumann, M. Sørensen, and A. Streller (1995). Stock returns and hyperbolic distributions. Technical report, Humboldt University of Berlin.
  - [13] Longin, F. M. (1996). The asymptotic distribution of extreme stock market returns. *Journal of Business* **69**, 383–408.
  - [14] Madan, D. B. and E. Seneta (1990). The variance gamma (V.G.) model for share market returns. *Journal of Business* **63**, 511–524.
  - [15] Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business* **36**, 394–419.
  - [16] Mandelbrot, B. (1967). The variation of some other speculative prices. *Journal of Business* **40**, 393–413.
  - [17] Mandelbrot, B. and H. M. Taylor (1967). On the distribution of stock price differences. *Operations Research* **15**, 1057–1062.
  - [18] Praetz, P. D. (1972). The distribution of share price changes. *Journal of Business* **45**, 49–55.
  - [19] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (2nd ed.). New York: Wiley.

# Some limit theory for $L_1$ -estimators in autoregressive models under general conditions

**Keith Knight**

*University of Toronto, Ontario, Canada*

*Abstract:* It is well-known that  $L_1$ -estimators of autoregressive parameters are asymptotically Normal if the distribution function of the errors,  $F(x)$ , has  $F'(0) = \lambda > 0$ . In this paper, we derive limiting distributions of  $L_1$ -estimators under more general assumptions on  $F$ . Second-order representations are also derived.

*Key words:* Autoregressive processes, limiting distributions, Bahadur-Kiefer theorems.

AMS subject classification: 62F12, 62M10, 60F05.

## 1 Introduction

$L_1$  estimation provides a somewhat robust alternative to least squares estimation for autoregressive models. Define a  $p$ -th order autoregressive (AR( $p$ )) process

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \quad (1)$$

where  $\{\varepsilon_t\}$  are independent, identically distributed (i.i.d.) random variables such that (a)  $E(\varepsilon_t^2) < \infty$ ; (b)  $\varepsilon_t$  has median 0; (c)  $F(x) = P(\varepsilon_t \leq x)$  is continuous at  $x = 0$ .

We will assume that the process  $\{Y_t\}$  is stationary; for this, we require that

$$\sum_{k=1}^p \phi_k z^k \neq 1$$

for all complex  $z$  with modulus  $|z| \leq 1$ . Throughout this paper, we will



assume the model (1) with the intercept  $\phi_0$ ; however, all of the results given in the paper will go through (with appropriate modifications) if  $\phi_0$  is suppressed and only  $\phi_1, \dots, \phi_p$  are estimated.

Least squares (or some related method) is typically used to estimate the parameters in the model (1). However, when the  $\varepsilon_t$ 's have heavy tails, least squares is inefficient compared to some other methods; one such method is  $L_1$ -estimation. We define  $L_1$ -estimators,  $\hat{\phi}_0, \hat{\phi}_1, \dots, \hat{\phi}_p$ , to minimize the objective function

$$g(v_0, v_1, \dots, v_p) = \sum_{t=1}^n |Y_t - v_0 - v_1 Y_{t-1} - \dots - v_p Y_{t-p}|. \quad (2)$$

(This assumes that we have  $n + p$  observations but asymptotically has no effect.) It is well-known (see Pollard, 1991; Wang and Wang, 1996) that the asymptotic behaviour of  $L_1$  estimators depends on the behaviour of the distribution function  $F(x)$  for  $x$  close to 0. For example, if  $F'(0) = \lambda > 0$  then we have

$$\sqrt{n}(\hat{\phi}_n - \phi) \rightarrow_d N_{p+1}(\mathbf{0}, C/(4\lambda^2)) \quad \text{as } n \rightarrow \infty$$

where  $C$  is a  $(p+1) \times (p+1)$  matrix defined to be

$$C = E[\mathbf{X}_t \mathbf{X}_t^T] \quad (3)$$

where  $\mathbf{X}_t = (1, Y_{t-1}, \dots, Y_{t-p})^T$ . Note that, contrary to popular belief, it is not necessary for  $F$  to be absolutely continuous to have asymptotic normality.

The assumption that  $F'(0) = \lambda > 0$  is quite strong in the sense that it is difficult to verify; given even very large samples, it is difficult to distinguish between a density which is finite at 0 and one which has a singularity at 0. For the sample median (which is the  $L_1$ -estimator of location), it has been shown that (for example, by de Haan and Taconis-Haantjes, 1979) that the rate of convergence depends on the behaviour of  $F(x)$  for  $x$  close to the population median; see also Smirnov (1952) who derives the domains of attraction for sample quantiles. Similarly, it is not necessary for  $F$  to be differentiable in order to find a limiting distribution for the  $L_1$ -estimator  $\hat{\phi}_n$ . We will assume that for some sequence  $\{a_n\}$  with  $a_n \rightarrow \infty$ , there exists a strictly increasing function  $\psi$  such that

$$\sqrt{n}(F(t/a_n) - F(0)) = \psi(t) + r_n(t) \quad (4)$$

where  $r_n(t) \rightarrow 0$  as  $n \rightarrow \infty$  for each  $t$ . Also define

$$\Psi(t) = \int_0^t \psi(s) ds \quad (5)$$

and

$$R_n(t) = \int_0^t r_n(s) ds. \quad (6)$$

Note that since  $\psi(t)$  is strictly increasing,  $\Psi(t)$  will be strictly convex.

The formulation given above provides a great deal of flexibility. For example, suppose that

$$F(x) - F(0) = \lambda \operatorname{sgn}(x) |x|^\alpha L(|x|)$$

for  $x$  in a neighbourhood of 0 where  $\alpha > 0$  and  $L$  is a slowly varying function at 0. ( $\operatorname{sgn}(x) = 1$  if  $x$  is positive and  $-1$  if  $x$  is negative.) In this case, we can take

$$a_n = n^{1/(2\alpha)} L^*(n) \quad \text{and} \quad \psi(t) = \lambda \operatorname{sgn}(t) |t|^\alpha$$

where  $L^*$  is a slowly varying function at infinity. When  $F(x)$  is differentiable at  $x = 0$  with  $F'(0) = \lambda > 0$  then  $\psi(t) = \lambda t$  and  $a_n = \sqrt{n}$ ; however, if  $F(x) - F(0) = \lambda x \ln(|x|^{-1})$  for  $x$  close to 0 then  $\psi(t) = \lambda t$  with  $a_n = \sqrt{n} \ln(n)/2$ . If  $F(x)$  is not differentiable at  $x = 0$  but has positive one-sided derivatives  $\lambda^+$  and  $\lambda^-$  then

$$\psi(t) = \begin{cases} \lambda^+ t & \text{for } t \geq 0 \\ \lambda^- t & \text{for } t < 0 \end{cases}.$$

(This occurs, for example, if the density has a jump at 0.)

In Section 2, we will determine the limiting distribution of the  $L_1$ -estimator under the general conditions on  $F$  described above, we will define

$$Z_n(\mathbf{u}) = \frac{a_n}{\sqrt{n}} \sum_{t=1}^n \left[ |\varepsilon_t - \mathbf{u}^T \mathbf{X}_t / a_n| - |\varepsilon_t| \right]. \quad (7)$$

Note that  $Z_n$  is minimized at  $\mathbf{u} = a_n(\hat{\phi}_n - \phi)$ .  $Z_n$  is a convex function and hence if the finite dimensional distributions converge weakly to those of a convex function  $Z$ , it follows that

$$a_n(\hat{\phi}_n - \phi) \rightarrow_d \operatorname{argmin}(Z)$$

provided  $\operatorname{argmin}(Z)$  is almost surely unique (Geyer, 1996). What is interesting is that only finite dimensional weak convergence is needed and not any sort of functional weak convergence (although this is implied by the finite dimensional convergence for convex functions).

In Section 3, we will obtain an “in distribution” Bahadur-Kiefer representation for the  $L_1$ -estimator under the general conditions described above. This will be done by approximating  $Z_n$  by an appropriate function  $Z_n^*$  and looking at the limiting behaviour of  $n^{1/4}(Z_n - Z_n^*)$ .

## 2 Limiting distributions

In order to derive the limiting distributions of the  $L_1$  estimators, we will assume the following regularity conditions.

(A1)  $\{\varepsilon_i\}$  are 0 median, finite variance i.i.d. random variables with distribution function  $F$  satisfying (4) for some  $\psi(t)$  and  $r_n(t)$ .

(A2) For each  $\mathbf{u}$ ,

$$E[\Psi(\mathbf{u}^T \mathbf{X}_t)] = \tau(\mathbf{u}) < \infty$$

where  $\Psi(t)$  is defined in (5) and  $\tau(\mathbf{u})$  is a strictly convex function.

(A3) For each  $\mathbf{u}$ ,

$$\frac{1}{n} \sum_{i=1}^n R_n(\mathbf{u}^T \mathbf{X}_t) \rightarrow_p 0$$

as  $n \rightarrow \infty$  where  $R_n(t)$  is defined in (6).

Note that condition (A1) implies that  $E[(\mathbf{u}^T \mathbf{X}_t)^2] < \infty$ ; thus, depending on the exact form of  $\Psi$ , condition (A2) may be implied by (A1). A sufficient condition for (A3) is  $E[|R_n(\mathbf{u}^T \mathbf{X}_t)|] \rightarrow 0$ .

**Theorem 1** *Suppose that  $\{Y_t\}$  is an  $AR(p)$  process satisfying (1) and that  $Z_n(\mathbf{u})$  is as defined in (7). Then under conditions (A1), (A2) and (A3),*

$$(Z_n(\mathbf{u}_1), \dots, Z_n(\mathbf{u}_k)) \rightarrow_d (Z(\mathbf{u}_1), \dots, Z(\mathbf{u}_k))$$

as  $n \rightarrow \infty$  where

$$Z(\mathbf{u}) = \mathbf{u}^T \mathbf{W} + 2\tau(\mathbf{u})$$

with  $\mathbf{W}$  a  $(p+1)$ -variate Normal random vector with mean  $\mathbf{0}$  and covariance matrix  $C$  defined in (3).

**Proof:** We will use the identity

$$|x - y| - |x| = y[I(x < 0) - I(x > 0)] + 2 \int_0^y [I(x \leq s) - I(x < 0)] ds$$

which is valid for  $x \neq 0$ . ( $I(A)$  is the indicator function of the set  $A$ .) Now

$$Z_n(\mathbf{u}) = Z_n^{(1)}(\mathbf{u}) + Z_n^{(2)}(\mathbf{u})$$

where

$$Z_n^{(1)}(\mathbf{u}) = \frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{X}_t^T \mathbf{u} [I(\varepsilon_t < 0) - I(\varepsilon_t > 0)]$$

$$\begin{aligned} \text{and } Z_n^{(2)}(\mathbf{u}) &= \frac{2a_n}{\sqrt{n}} \sum_{t=1}^n \int_0^{v_{nt}} [I(\varepsilon_t \leq s) - I(\varepsilon_t \leq 0)] ds \\ &= \sum_{t=1}^n Z_{nt}^{(2)}(\mathbf{u}) \end{aligned}$$

(with  $v_{nt} = \mathbf{X}_t^T \mathbf{u} / a_n$ ). Since, for each  $\mathbf{u}$ , the summands in  $Z_n^{(1)}(\mathbf{u})$  are stationary martingale differences with finite variance, it follows from a martingale central limit theorem that

$$Z_n^{(1)}(\mathbf{u}) \rightarrow_d \mathbf{u}^T \mathbf{W}$$

and the convergence in distribution holds for any finite collection of  $\mathbf{u}$ 's. For  $Z_n^{(2)}(\mathbf{u})$ , we have

$$Z_n^{(2)}(\mathbf{u}) = \sum_{t=1}^n E(Z_{nt}^{(2)}(\mathbf{u})) + \sum_{t=1}^n (Z_{nt}^{(2)}(\mathbf{u}) - E(Z_{nt}^{(2)}(\mathbf{u}))).$$

Letting  $v_t = \mathbf{X}_t^T \mathbf{u} = a_n v_{nt}$ , it follows that

$$\begin{aligned} \sum_{t=1}^n E(Z_{nt}^{(2)}(\mathbf{u})) &= \frac{2a_n}{\sqrt{n}} \sum_{t=1}^n \int_0^{v_{nt}} (F(s) - F(0)) ds \\ &= \frac{2}{n} \sum_{t=1}^n \int_0^{v_t} \sqrt{n} (F(s/a_n) - F(0)) ds \\ &= \frac{2}{n} \sum_{t=1}^n [\Psi(\mathbf{u}^T \mathbf{X}_t) + R_n(\mathbf{u}^T \mathbf{X}_t)] \\ &\rightarrow 2\tau(\mathbf{u}) \end{aligned}$$

where  $R_n$  is defined in (6). For the remainder term in  $Z_n^{(2)}(\mathbf{u})$ , we have (since the summands are again martingale differences)

$$\begin{aligned} \text{Var}(Z_n^{(2)}(\mathbf{u})) &= \sum_{t=1}^n E[(Z_{nt}^{(2)}(\mathbf{u}) - E(Z_{nt}^{(2)}(\mathbf{u})))^2] \\ &\leq \frac{2}{\sqrt{n}} \max_{1 \leq t \leq n} |\mathbf{X}_t^T \mathbf{u}| \sum_{t=1}^n E(Z_{nt}^{(2)}(\mathbf{u})) \\ &= \frac{2}{\sqrt{n}} \max_{1 \leq t \leq n} |\mathbf{X}_t^T \mathbf{u}| E(Z_n^{(2)}(\mathbf{u})). \end{aligned}$$

$\{\mathbf{X}_t^T \mathbf{u}\}$  is stationary with finite second moment and so

$$\max_{1 \leq t \leq n} |\mathbf{X}_t^T \mathbf{u}| \rightarrow_p 0.$$

Thus

$$Z_n^{(2)}(\mathbf{u}) - E(Z_n^{(2)}(\mathbf{u})) \rightarrow_p 0 \quad \text{as } n \rightarrow \infty$$

and so  $Z_n^{(2)}(\mathbf{u}) \rightarrow_p 2\tau(\mathbf{u})$ . Thus we have

$$Z_n(\mathbf{u}) \rightarrow_d \mathbf{u}^T \mathbf{W} + 2\tau(\mathbf{u}) = Z(\mathbf{u})$$

and the finite dimensional convergence holds trivially.  $\square$

The following corollary gives us a representation of the limiting distribution of  $a_n(\hat{\phi}_n - \phi)$ .

**Corollary 2** *Let  $\hat{\phi}_n$  minimize (2). Under the assumptions of Theorem 1,*

$$a_n(\hat{\phi}_n - \phi) \rightarrow_d \operatorname{argmin}(Z)$$

as  $n \rightarrow \infty$ .

**Proof:** Since  $\tau$  is strictly convex,  $Z_\alpha$  is strictly convex and so has a unique minimum. The result follows from Geyer (1996).  $\square$

The limiting distribution given in Corollary 2 will not be normal unless the function  $\tau(\mathbf{u})$  is quadratic. In the following example, we illustrate the computation of the limiting distribution in a special case.

**Example 1** Consider the AR(1) process

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t$$

where the  $\varepsilon_t$ 's are i.i.d. random variables with density

$$f_\alpha(x) = \frac{|x|^{\alpha-1} \exp(-|x|)}{2\Gamma(\alpha)}$$

for some  $\alpha > 0$ . (This is a two-sided Gamma distribution.) For  $x$  close to 0, we have

$$F_\alpha(x) - F_\alpha(0) \approx \frac{\operatorname{sgn}(x)|x|^\alpha \exp(-x)}{2\Gamma(\alpha+1)}$$

and so setting  $a_n = n^{1/(2\alpha)}$ , we get

$$\sqrt{n}(F_\alpha(t/a_n) - F_\alpha(0)) \rightarrow \psi_\alpha(t) = \frac{1}{2\Gamma(\alpha+1)} \operatorname{sgn}(t)|t|^\alpha$$

with

$$|\sqrt{n}(F_\alpha(t/a_n) - F_\alpha(0)) - \psi_\alpha(t)| \leq k(\alpha) \frac{|t|^{\alpha+1}}{n^{1/(2\alpha)}}.$$

It is easy to verify that conditions (A1), (A2) and (A3) are all satisfied (since  $E[|Y_t|^r] < \infty$  for all  $r > 0$ ) and so for given  $\phi_0, \phi_1$ , the limiting objective function in Theorem 1 is

$$Z_\alpha(u_0, u_1) = u_0 W_0 + u_1 W_1 + \frac{1}{\Gamma(\alpha + 2)} E[|u_0 + u_1 Y_1|^{\alpha+1}]$$

where  $W_0$  and  $W_1$  are zero mean Normal random variables with  $\text{Var}(W_0) = 1$ ,  $\text{Var}(W_1) = E(Y_1^2)$  and  $\text{Cov}(W_0, W_1) = E(Y_1)$ . By differentiation, we determine the minimizers of  $Z_\alpha, \hat{U}_0$  and  $\hat{U}_1$ , to satisfy the equations

$$\begin{aligned} W_0 + \frac{1}{\Gamma(\alpha + 1)} d_0(\hat{U}_0, \hat{U}_1) &= 0 \\ W_1 + \frac{1}{\Gamma(\alpha + 1)} d_1(\hat{U}_0, \hat{U}_1) &= 0 \end{aligned}$$

where

$$\begin{aligned} d_0(u_0, u_1) &= E[\text{sgn}(u_0 + u_1 Y_1) |u_0 + u_1 Y_1|^\alpha] \quad \text{and} \\ d_1(u_0, u_1) &= E[\text{sgn}(u_0 + u_1 Y_1) Y_1 |u_0 + u_1 Y_1|^\alpha]. \end{aligned}$$

( $\text{sgn}(x) = 1$  or  $-1$  depending on whether  $x$  is positive or negative.)

If  $f_W(w_0, w_1)$  is the joint density of  $(W_0, W_1)$ , it then follows that the joint density of  $(\hat{U}_0, \hat{U}_1)$  (that is, the limiting density of  $a_n(\hat{\phi}_n - \phi)$ ) is

$$f_U(u_0, u_1) = \frac{1}{\Gamma(\alpha)^2} f_W \left( \frac{d_0(u_0, u_1)}{\Gamma(\alpha + 1)}, \frac{d_1(u_0, u_1)}{\Gamma(\alpha + 1)} \right) (d_{00}(u_0, u_1) d_{11}(u_0, u_1) - d_{10}^2(u_0, u_1))$$

where

$$\begin{aligned} d_{00}(u_0, u_1) &= E[|u_0 + u_1 Y_1|^{\alpha-1}] \\ d_{11}(u_0, u_1) &= E[Y_1^2 |u_0 + u_1 Y_1|^{\alpha-1}] \quad \text{and} \\ d_{10}(u_0, u_1) &= E[Y_1 |u_0 + u_1 Y_1|^{\alpha-1}]. \end{aligned}$$

The density  $f_U$  cannot easily be computed analytically (unless  $\phi_1 = 0$ ) but can be computed feasibly using Monte Carlo sampling.

### 3 Second order properties

It follows from the proof of Theorem 1 that we can approximate  $Z_n$  by the function

$$Z_n^*(\mathbf{u}) = -\frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{X}_t^T \mathbf{u} [I(\varepsilon_t > 0) - I(\varepsilon_t < 0)] + 2\tau(\mathbf{u}). \quad (8)$$

It is easy to see that we can approximate  $a_n(\hat{\phi}_n - \phi)$  by the minimizer of  $Z_n^*$ . For example, if  $\psi(t) = \lambda t$  (for some  $\lambda > 0$ ) and  $a_n = \sqrt{n}$ , it follows that  $\tau(\mathbf{u}) = \lambda \mathbf{u}^T C \mathbf{u} / 2$  and so we can approximate  $\sqrt{n}(\hat{\phi}_n - \phi)$  by

$$\frac{1}{2\lambda\sqrt{n}} \sum_{t=1}^n C^{-1} \mathbf{X}_t [I(\varepsilon_t > 0) - I(\varepsilon_t < 0)].$$

More generally, we have

$$a_n(\hat{\phi}_n - \phi) \approx \mathbf{h}^{-1}(\mathbf{W}_n/2)$$

where  $\mathbf{h}(\mathbf{u})$  is the gradient of  $\tau(\mathbf{u})$ ,  $\mathbf{h}^{-1}$  its inverse and

$$\mathbf{W}_n = \frac{1}{\sqrt{n}} \sum_{t=1}^n \mathbf{X}_t [I(\varepsilon_t > 0) - I(\varepsilon_t < 0)]. \quad (9)$$

(Typically,  $\mathbf{h}(\mathbf{u}) = E[\mathbf{X}_t \psi(\mathbf{u}^T \mathbf{X}_t)]$ .) Theorems which deal with the asymptotic behaviour of this approximation error are commonly known as Bahadur-Kiefer theorems due to their connection with the work of Bahadur (1966) and Kiefer (1967) for sample quantiles. What will be proved below is an “in distribution” (as opposed to “almost sure”) Bahadur-Kiefer theorem. The following lemma will be useful in determining the asymptotic behaviour of the approximation error.

**Lemma 3** *Define*

$$g_n(\mathbf{u}) = -\mathbf{x}_n^T \mathbf{u} + \rho_n(\mathbf{u}) \quad h_n(\mathbf{u}) = -\mathbf{x}_n^T \mathbf{u} + \rho(\mathbf{u})$$

and let  $\mathbf{u}_n = \operatorname{argmin}(g_n)$ ,  $\mathbf{v}_n = \operatorname{argmin}(h_n)$ . Suppose that

- (i)  $\mathbf{x}_n \rightarrow \mathbf{x}_0$ ;
- (ii)  $\mathbf{u}_n - \mathbf{v}_n \rightarrow 0$ ;
- (iii) for any  $t$ ,  $\mathbf{u}$  and  $\mathbf{w}$ ,

$$\rho_n(\mathbf{u} + t\mathbf{w}) - \rho_n(\mathbf{u}) = \int_0^t \mathbf{w}^T \psi_n(\mathbf{u} + s\mathbf{w}) ds$$

and

$$\rho(\mathbf{u} + t\mathbf{w}) - \rho(\mathbf{u}) = \int_0^t \mathbf{w}^T \psi(\mathbf{u} + s\mathbf{w}) ds$$

for some functions  $\{\psi_n\}$  and  $\psi$  where  $\psi$  is one-to-one.

- (iv)  $\mathbf{v}_0 = \psi^{-1}(\mathbf{x}_0)$  exists and for some  $\alpha > 0$

$$\|\psi(\mathbf{u}) - \psi(\mathbf{v})\| \leq k \|\mathbf{u} - \mathbf{v}\|^\alpha$$

for all  $\mathbf{u}, \mathbf{v}$  in a neighbourhood of  $\mathbf{v}_0$ ;

(v) For some sequence  $\{b_n\}$  with  $b_n \rightarrow \infty$  and any compact set  $K$ ,

$$\sup_{\mathbf{u} \in K} \|b_n(\psi_n(\mathbf{u}) - \psi(\mathbf{u})) - \mathbf{d}_0(\mathbf{u})\| \rightarrow 0$$

where  $\mathbf{d}_0$  is a continuous function.

Then

$$b_n(\psi(\mathbf{u}_n) - \psi(\mathbf{v}_n)) \rightarrow -\mathbf{d}_0(\mathbf{v}_0)$$

where  $\mathbf{v}_0 = \psi^{-1}(\mathbf{x}_0)$ .

A proof of Lemma 3 will not be given here. Note that if the function  $\psi(\mathbf{u})$  has continuous partial derivatives at  $\mathbf{u} = \mathbf{v}_0$  then under the conditions of Lemma 3 we have

$$b_n(\mathbf{u}_n - \mathbf{v}_n) \rightarrow -H_\rho^{-1}(\mathbf{v}_0)\mathbf{d}_0(\mathbf{v}_0)$$

provided  $H_\rho(\mathbf{u})$ , the Hessian of  $\rho$ , is invertible at  $\mathbf{u} = \mathbf{v}_0$ .

Lemma 3 will be applied to sequences of random elements by appealing to a Skorokhod-type arguments (see, for example, van der Vaart and Wellner, 1996) to construct almost surely convergent sequences. To do this, we will define a space  $B_r(R^d)$  of locally bounded  $R^r$ -valued functions defined on  $R^d$ . (By “locally bounded” we mean bounded on compact sets.) If  $\{\mathbf{g}_n\}$  and  $\mathbf{g}$  are elements of  $B_r(R^d)$  then we will say that  $\{\mathbf{g}_n\}$  converges to  $\mathbf{g}$  if

$$\sup_{\mathbf{u} \in K} \|\mathbf{g}_n(\mathbf{u}) - \mathbf{g}(\mathbf{u})\| \rightarrow 0$$

for all compact subsets  $K$  of  $R^d$ . A possible metric for this topology is

$$d(\mathbf{g}, \mathbf{h}) = \sum_{k=1}^{\infty} \min(1, d_k(\mathbf{g}, \mathbf{h}))2^{-k}$$

where

$$d_k(\mathbf{g}, \mathbf{h}) = \sup_{\|\mathbf{u}\| \leq k} \|\mathbf{g}(\mathbf{u}) - \mathbf{h}(\mathbf{u})\|.$$

We also define  $C_r(R^d)$  to be the space of  $R^r$ -valued continuous functions on  $R^d$ ;  $C_r(R^d)$  is a separable subset of  $B_r(R^d)$ . If  $\{\mathbf{D}_n\}$  and  $\mathbf{D}$  are random elements of  $B_r(R^d)$  such that  $\mathbf{D}_n \rightarrow_d \mathbf{D}$  and  $\mathbf{D}$  is (with probability 1) a random element of  $C_r(R^d)$  then it is possible find almost surely convergent representations of  $\{\mathbf{D}_n\}$  and  $\mathbf{D}$ .



(A4) For each compact set  $K$ ,

$$\sup_{u \in K} n^{1/4} \left\| E[\mathbf{X}_t r_n(\mathbf{u}^T \mathbf{X}_t)] \right\| \rightarrow 0$$

as  $n \rightarrow \infty$ .

(A5) For each compact set  $K$ , we have

$$\sup_{u \in K} \left| \frac{1}{n} \sum_{t=1}^n \mathbf{X}_t^T \mathbf{X}_t r_n(\mathbf{u}^T \mathbf{X}_t) \right| \rightarrow_p 0$$

as  $n \rightarrow \infty$ .

(A6) For each  $\mathbf{u}$ ,  $E[\mathbf{X}_t^T \mathbf{X}_t |\psi(\mathbf{u}^T \mathbf{X}_t)|]$  is finite.

**Theorem 4** Suppose that  $\{Y_t\}$  is an  $AR(p)$  process satisfying (1) with  $Z_n$  and  $Z_n^*$  defined as in (7) and (8). Then under conditions (A1)-(A6), we have

$$n^{1/4}(Z_n(\mathbf{u}) - Z_n^*(\mathbf{u})) \rightarrow_d V(\mathbf{u}) \quad \text{as } n \rightarrow \infty$$

on  $C_1(R^{p+1})$  where

$$V(\mathbf{u} + t\mathbf{w}) - V(\mathbf{u}) = 2 \int_0^t \mathbf{w}^T \mathbf{D}(\mathbf{u} + s\mathbf{w}) ds$$

and  $\mathbf{D}(\mathbf{u})$  is a zero mean Gaussian process with  $\mathbf{D}(\mathbf{0}) = \mathbf{0}$  and

$$E[(\mathbf{D}(\mathbf{u}) - \mathbf{D}(\mathbf{v}))(\mathbf{D}(\mathbf{u}) - \mathbf{D}(\mathbf{v}))^T] = E[\mathbf{X}_t \mathbf{X}_t^T |\psi(\mathbf{u}^T \mathbf{X}_t) - \psi(\mathbf{v}^T \mathbf{X}_t)|].$$

**Proof:** Define

$$V_n(\mathbf{u}) = n^{1/4}(Z_n(\mathbf{u}) - Z_n^*(\mathbf{u}))$$

and note that  $V_n(\mathbf{0}) = 0$  for all  $n$ . We also have

$$V_n(\mathbf{u} + t\mathbf{w}) - V_n(\mathbf{u}) = 2 \int_0^t \mathbf{w}^T \mathbf{D}_n(\mathbf{u} + s\mathbf{w}) ds$$

where

$$\begin{aligned} \mathbf{D}_n(\mathbf{u}) = & \frac{1}{n^{1/2}} \sum_{t=1}^n [n^{1/4} \mathbf{X}_t (I(\varepsilon_t \leq \mathbf{u}^T \mathbf{X}_t / a_n) \\ & - I(\varepsilon_t \leq 0)) - n^{-1/4} E[\mathbf{X}_t \psi(\mathbf{u}^T \mathbf{X}_t)]] \end{aligned}$$

since our assumptions imply that the gradient of  $\tau(\mathbf{u}) = E[\Psi(\mathbf{u}^T \mathbf{X}_t)]$  is  $E[\mathbf{X}_t \psi(\mathbf{u}^T \mathbf{X}_t)]$ . Clearly,  $\mathbf{D}_n(\mathbf{0}) = \mathbf{0}$  for all  $n$  and applying an appropriate martingale central limit theorem (Hall and Heyde, 1980), it follows that

the finite dimensional distributions of  $\mathbf{D}_n$  converge to those of  $\mathbf{D}$ . It is also straightforward to verify that on each compact set  $K$ ,

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} P \left( \sup_{\|u-v\| < \delta; u, v \in K} \|\mathbf{D}_n(\mathbf{u}) - \mathbf{D}_n(\mathbf{v})\| > \epsilon \right) = 0$$

for every  $\epsilon > 0$  by using an appropriate moment condition. Hence  $\mathbf{D}_n \rightarrow_d \mathbf{D}$  on  $B_{p+1}(R^{p+1})$  and the conclusion follows.  $\square$

**Theorem 5** *Assume the conditions of Theorem 4 and let  $\mathbf{h}(\mathbf{u})$  be the gradient of  $\tau(\mathbf{u})$  with inverse  $\mathbf{h}^{-1}$ . If  $\mathbf{U}$  minimizes  $Z$  and*

$$\|\mathbf{h}(\mathbf{u}) - \mathbf{h}(\mathbf{v})\| \leq k \|\mathbf{u} - \mathbf{v}\|^\alpha \quad (\alpha > 0)$$

for all  $\mathbf{u}, \mathbf{v}$  in a neighbourhood of  $\mathbf{U}$  ( $k$  and  $\alpha$  may depend on  $\mathbf{U}$ ) then

$$n^{1/4} \left( \mathbf{h}(a_n(\hat{\phi}_n - \phi)) - \frac{\mathbf{W}_n}{2} \right) \rightarrow_d -\mathbf{D}(\mathbf{h}^{-1}(\mathbf{W}/2))$$

as  $n \rightarrow \infty$  where  $\mathbf{W}_n$  is defined in (9),  $\mathbf{D}$  is the Gaussian process defined in Theorem 4 and  $\mathbf{W}$  is a  $(p+1)$ -variate Normal random vector (independent of  $\mathbf{D}$ ) with mean  $\mathbf{0}$  and covariance matrix  $\mathbf{C}$ .

**Proof:** Let  $\mathbf{U}_n = a_n(\hat{\phi}_n - \phi)$  minimize  $Z_n$ . Then it is easy to verify that

$$(\mathbf{U}_n, n^{1/4}(Z_n - Z_n^*)) \rightarrow_d (\mathbf{h}^{-1}(\mathbf{W}/2), V)$$

as  $n \rightarrow \infty$  on the space  $R^{p+1} \times B_1(R^{p+1})$  where  $\mathbf{W}$  and  $V$  are independent. Since the limit is concentrated on a separable subset of  $R^{p+1} \times B_1(R^{p+1})$  (namely  $R^{p+1} \times C_1(R^{p+1})$ ), we can construct a probability space and almost surely convergent versions of  $\{\mathbf{U}_n\}$  and  $\{n^{1/4}(Z_n - Z_n^*)\}$ . The conclusion follows by applying Lemma 3 to each convergent sequence.  $\square$

If  $\mathbf{h}(\mathbf{u})$  is one-to-one (with inverse  $\mathbf{h}^{-1}$ ) and continuously differentiable then it follows that (under the conditions of Theorem 5)

$$n^{1/4} \left[ a_n(\hat{\phi}_n - \phi) - \mathbf{h}^{-1}(\mathbf{W}_n/2) \right] \rightarrow_d -H^{-1}(\mathbf{h}^{-1}(\mathbf{W}/2))\mathbf{D}(\mathbf{h}^{-1}(\mathbf{W}/2))$$

provided that  $H(\mathbf{u})$ , the Hessian of  $\tau$ , is invertible outside of a set of Lebesgue measure 0 in  $R^{p+1}$ . This suggests the asymptotic expansion

$$\begin{aligned} a_n(\hat{\phi}_n - \phi) &= \mathbf{h}^{-1}(\mathbf{W}_n/2) \\ &\quad - \frac{1}{n^{1/4}} H^{-1}(\mathbf{h}^{-1}(\mathbf{W}_n/2))\mathbf{D}(\mathbf{h}^{-1}(\mathbf{W}_n/2)) + o_p(n^{-1/4}). \end{aligned}$$

(As before,  $\mathbf{W}_n$  is defined as in (9).) Whether this expansion is particularly useful is an open question.

Evaluating the limiting distribution in Theorem 5 is tedious but not overly difficult (provided, of course, that everything about  $\{Y_t\}$  is known). For a given  $\mathbf{u}$ ,  $\mathbf{D}(\mathbf{u})$  is  $(p+1)$ -variate Normal with mean  $\mathbf{0}$  and covariance matrix

$$K(\mathbf{u}) = E[\mathbf{X}_t \mathbf{X}_t^T | \psi(\mathbf{u}^T \mathbf{X}_t)].$$

If  $K(\mathbf{u})$  is positive definite for  $\mathbf{u}$  outside of a set of Lebesgue measure 0 then since  $\mathbf{W}$  is independent of  $\mathbf{D}$ , it follows that the density of  $-\mathbf{D}(\mathbf{h}^{-1}(\mathbf{W}/2))$  is

$$f_1(\mathbf{x}) = \frac{1}{|C|^{1/2} \pi^{p+1}} \int \frac{|H(\mathbf{u})|}{|K(\mathbf{u})|^{1/2}} \exp \left[ -\frac{1}{2} \gamma_1(\mathbf{x}, \mathbf{u}) \right] d\mathbf{u}$$

where

$$\gamma_1(\mathbf{x}, \mathbf{u}) = \mathbf{x}^T K^{-1}(\mathbf{u}) \mathbf{x} + 4\mathbf{h}(\mathbf{u})^T C^{-1} \mathbf{h}(\mathbf{u})$$

and the integration is over  $R^{p+1}$  with  $|\cdot|$  denoting determinant. Likewise the density of  $-H^{-1}(\mathbf{h}^{-1}(\mathbf{W}/2))\mathbf{D}(\mathbf{h}^{-1}(\mathbf{W}/2))$  is

$$f_2(\mathbf{x}) = \frac{1}{|C|^{1/2} \pi^{p+1}} \int \frac{|H(\mathbf{u})|^2}{|K(\mathbf{u})|^{1/2}} \exp \left[ -\frac{1}{2} \gamma_2(\mathbf{x}, \mathbf{u}) \right] d\mathbf{u}$$

where

$$\gamma_2(\mathbf{x}, \mathbf{u}) = \mathbf{x}^T H(\mathbf{u}) K^{-1}(\mathbf{u}) H(\mathbf{u}) \mathbf{x} + 4\mathbf{h}(\mathbf{u})^T C^{-1} \mathbf{h}(\mathbf{u}).$$

In the following example, we derive the density  $f_2$  in a simple case.

**Example 2.** Let  $Y_t = \varepsilon_t$  where  $\{\varepsilon_t\}$  are i.i.d. random variables with distribution function  $F$  satisfying  $F'(0) = \lambda > 0$ . Suppose that we estimate only the parameter  $\phi_1$  of an AR(1) model; call this estimator  $\hat{\phi}_n$ . We then have

$$\tau(u) = \frac{1}{2} \lambda \sigma^2 u^2$$

where  $\sigma^2 = E(\varepsilon_t^2)$ . We also have  $C = \sigma^2$  and  $H(u) = \tau''(u) = \lambda \sigma^2$ . Finally,  $K(u) = \lambda \gamma |u|$  where  $\gamma = E[|\varepsilon_t|^3]$ . It follows from Theorems 4 and 5 that

$$n^{1/4} \left( \sqrt{n} \hat{\phi}_n - \frac{1}{2\lambda\sqrt{n}} \sum_{t=1}^n Y_{t-1} [I(\varepsilon_t > 0) - I(\varepsilon_t < 0)] \right) \rightarrow_d S$$

where  $S$  has density

$$f_S(x) = \frac{\lambda^{3/2} \sigma^2}{\pi \gamma^{1/2}} \int_{-\infty}^{\infty} |u|^{-1/2} \exp \left[ -\frac{1}{2} \left( \frac{\lambda \sigma^4 x^2}{\gamma |u|} + 4\lambda^2 \sigma^2 u^2 \right) \right] du.$$

$S$  has a symmetric distribution with moment generating function

$$E[\exp(tS)] = 2 \exp\left(\frac{\gamma^2 t^4}{32\lambda^4 \sigma^{10}}\right) \Phi\left(\frac{\gamma t^2}{2\lambda^2 \sigma^5}\right)$$

where  $\Phi$  is the standard Normal distribution function.

## 4 Final comments

In this paper, we have derived first- and second-order limiting distributions for the  $L_1$ -estimators of the parameters of an  $\text{AR}(p)$  process under fairly general conditions on the error distribution. From a statistical point of view, the fact that the asymptotic behaviour of the  $L_1$ -estimators is so sensitive to the behaviour of  $F(x)$  for  $x$  close to 0 is somewhat troubling. One possible non-parametric approach to estimating the sampling distribution of  $\hat{\phi}_n$  is to bootstrap the  $\text{AR}(p)$  process by sampling with replacement from the residuals  $e_t = Y_t - \mathbf{X}_t^T \hat{\phi}_n$ . However, it is possible to show that, asymptotically, this bootstrap procedure is correct to first order only if  $\psi(t)$  is a linear function and is never correct to second order. This is similar to the results of Hall and Martin (1988) and Huang *et al.* (1996) for sample quantiles of i.i.d. random variables. However, other approaches to bootstrapping time series, such as frequency domain bootstrapping, may prove to be more fruitful in this problem.

It may also be possible to exploit Lemma 3 to obtain an “almost sure” Bahadur-Kiefer representation. Arcones (1996a, 1996b) and He and Shao (1996) derive such representations for  $L_p$ -estimators in linear regression models. However, these papers assume that  $F(x)$ , the distribution function of the errors, is linear in a neighbourhood of  $x = 0$ . Using the notation of Theorems 4 and 5, we can conjecture that, under appropriate regularity conditions,

$$(n/\ln(\ln(n)))^{1/4} \left( h(b_n(\hat{\phi}_n - \phi)) - \frac{\mathbf{W}_n}{2\sqrt{\ln(\ln(n))}} \right) = O(1) \quad (10)$$

with probability 1 where  $b_n$  satisfies the condition

$$\lim_{n \rightarrow \infty} \sqrt{n/\ln(\ln(n))} (F(t/b_n) - F(0)) = \psi(t)$$

and the set of limit points of the left hand side of (10) is non-trivial.

## Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author would like to thank a referee for some useful comments.

## References

- [1] Arcones, M.A. (1996a). The Bahadur-Kiefer representation of  $L_p$  regression estimators. *Econometric Theory* **12**, 257-283.
- [2] Arcones, M.A. (1996b). Second order representations of the least absolute deviation regression estimator. Unpublished manuscript.
- [3] Bahadur, R.R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37**, 577-580.
- [4] de Haan, L. and E. Taconis-Haantjes (1979). On Bahadur's representation of sample quantiles. *Ann. Inst. Statist. Math.* **31**, 299-308.
- [5] Geyer, C.J. (1996). On the asymptotics of convex stochastic optimization. Unpublished manuscript.
- [6] Hall, P. and C.C. Heyde (1980). *Martingale Limit Theory and its Application*. New York: Academic Press.
- [7] Hall, P. and M.A. Martin (1988). Exact convergence rate of bootstrap quantile variance estimator. *Probability theory and related fields* **80**, 261-268.
- [8] He, X. and Q.-M. Shao (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24**, 2608-2631.
- [9] Huang, J.S., P.K. Sen and J. Shao (1996). Bootstrapping a sample quantile when the density has a jump. *Statistica Sinica* **6**, 299-309.
- [10] Kiefer, J. (1967). On Bahadur's representation of sample quantiles. *Ann. Math. Statist.* **38**, 1323-1342.
- [11] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7**, 186-199.
- [12] Smirnov, N.V. (1952). Limit distributions for the terms of a variational series. *American Mathematical Society Translation Series (1)* **11**, 82-143.
- [13] van der Vaart, A.W. and J.A. Wellner (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. New York: Springer.
- [14] Wang, L. and J. Wang (1996). Limit distribution of  $L_1$ -estimators for time series: stationary linear models. Unpublished manuscript.

# A Lagrange multiplier approach to testing for serially dependent error terms

Hans Nyquist

*University of Umeå, Sweden*

*Abstract:* Testing for first-order auto-regressive errors in a linear regression model is considered. It is found that the  $L_1$ -norm based Lagrange multiplier test avoids computational difficulties caused by the dependency among the errors. Furthermore, the Lagrange multiplier test has the advantage that estimation of the error term density at zero is not required. As the error term variance increases and the error term density at zero becomes larger the asymptotic relative efficiency becomes more favorable for the  $L_1$ -norm based test relative to the corresponding least squares test.

*Key words:* Auto-regressive errors,  $L_1$ -norm estimation, Lagrange multiplier test.

AMS subject classification: 62J05, 62M10.

## 1 Introduction

The existence of serially dependent error terms when a linear regression model is used for analyzing time series data has attracted the attention of recent research. Elimination of estimation problems caused by the dependence involves both the detection of the dependence and the selection of an appropriate estimation technique for the model in question if dependence is found.

It is also well recognized in the literature that many data sets contain outliers or, alternatively, are well represented by distributions with fat tails. This has motivated the introduction of robust estimators, including the  $L_1$ -norm estimator. Simulation experiments of  $L_1$ -norm based estimators of models with serially dependent error terms are reported in e.g. Coursey and Nyquist (1983 and 1986). These experiments indicate that  $L_1$ -norm

methods for estimation are preferable to least squares methods when the tails of the error distribution becomes fatter. The experiments also indicate that estimators that take account of the serially dependence outperform those estimators that ignore it.

Modelling for serially dependent errors results in a non-linear model, implying that standard techniques for computing estimates do not apply. One procedure for least squares estimation of linear regression models with serially dependent errors is that of Cochrane and Orcutt (1949). The Cochrane-Orcutt approach was extended to estimation using a more general criterion function in Coursey and Nyquist (1986) and further analyzed in Nyquist (1992). These analyses show that saddle points and multiple minima are very common when the  $L_1$ -norm criterion function is used. As a consequence, the  $L_1$ -norm based Cochrane-Orcutt procedure has a tendency to converge to points that do not define the global minimum of the criterion function. On the other hand, extensions of the Gauss-Newton technique to the  $L_1$ -norm case, as suggested by Osborne and Watson (1971) and Anderson and Osborne (1977), has shown more satisfactory results when it has been applied to models with serially dependent errors (Nyquist, 1992).

In this paper attention is restricted to the case with a first-order autoregressive error process. The aim of the paper is to present and discuss the Lagrange multiplier test of the hypothesis of serially independent error terms. This test is particularly attractive from a computational viewpoint since it only requires estimation of the restricted model, which is in this case, a linear model. Computational difficulties caused by non-linearities are therefore avoided in this approach. Furthermore, the Lagrange multiplier test has the advantage that it does not require estimation of the error density at zero.

## 2 The model

We consider the linear regression model with first-order auto-regressive errors

$$y_t = x_t' \beta + u_t, \quad t = 1, 2, \dots, T, \quad (1)$$

$$u_t = \phi u_{t-1} + v_t, \quad t = 1, 2, \dots, T, \quad (2)$$

where  $y_t$  is a response variable,  $x_t$  is a  $p$ -vector of known regressors,  $\beta$  is a  $p$ -vector of unknown regression parameters, and  $u_t$  is the error generated by the auto-regressive process with parameter  $\phi$ . The disturbances  $v_1, v_2, \dots, v_T$  are assumed to be independent and identically distributed.

Let  $X$  be the  $T \times p$  matrix with  $x'_t$  as rows. For the subsequent analysis we assume that

*Assumption 1:* There exists a positive definite matrix  $D$  such that  $\lim_{T \rightarrow \infty} T^{-1} X' X = D$ .

*Assumption 2:* The common cumulative distribution function of the disturbances  $v_t$ ,  $F$ , is differentiable at 0 with  $F'(0) > 0$ .

*Assumption 3:* The disturbances  $v_t$  have a finite variance,  $d = V(v_t)$ .

Our aim is to test the null hypothesis

$$H_0 : \phi = 0,$$

against the alternative

$$H_a : \phi \neq 0.$$

Lagging (1) by one period, multiplying by  $\phi$ , and subtracting from (1) yields

$$y_t - \phi y_{t-1} = (x'_t - \phi x'_{t-1}) \beta + v_t, \quad t = 2, 3, \dots, T.$$

Conditioning on the first observation, the  $L_1$ -norm criterion function is defined as

$$\begin{aligned} S(\beta, \phi) &= \sum_{t=2}^T |y_t - \phi y_{t-1} - (x'_t - \phi x'_{t-1}) \beta| \\ &= \sum_{t=2}^T |y_t - h_t(\beta, \phi)|, \end{aligned} \quad (3)$$

where

$$h_t(\beta, \phi) = \phi y_{t-1} + (x'_t - \phi x'_{t-1}) \beta. \quad (4)$$

The  $L_1$ -norm estimator  $(\hat{\beta}, \hat{\phi})$  is defined as the minimizer of  $S(\beta, \phi)$ . The computational problems caused by the serially dependency among the error terms stem from the non-linearity of  $h_t(\beta, \phi)$ .

The Gauss-Newton approach to minimize  $S(\beta, \phi)$  utilizes a first-order Taylor approximation of  $h_t(\beta, \phi)$  evaluated at a previous estimate  $(\beta^{(s)}, \phi^{(s)})$  of the parameters,



$$h_t^{(s)}(\beta, \phi) =$$

$$h_t(\beta^{(s)}, \phi^{(s)}) + (\phi - \phi^{(s)}) \hat{u}_{t-1}^{(s)} + (x'_t - \phi^{(s)} x'_{t-1}) (\beta - \beta^{(s)}), \quad (5)$$

where

$$\hat{u}_t^{(s)} = y_t - x'_t \beta^{(s)}.$$

The minimizer of

$$S^{(s)}(\beta, \phi) = \sum_{t=2}^T |y_t - h_t^{(s)}(\beta, \phi)| \quad (6)$$

now yields new estimates,  $(\beta^{(s+1)}, \phi^{(s+1)})$ . The function  $h_t^{(s)}(\beta, \phi)$  is linear in the parameters, so that standard routines, such as those of Barrodale and Roberts (1974) and Armstrong et al. (1979), can be used for finding the minimum of (6). Conditions for the convergence of the iteration process are found in Osborne and Watson (1971) and Anderson and Osborne (1977).

### 3 The Lagrange multiplier test

The Lagrange multiplier test is based on the gradient of the unrestricted estimation problem evaluated at the restricted estimate. Thus, under the null hypothesis that  $\phi = 0$ , the restricted model is the linear regression model ignoring dependent error terms. Denoting the restricted estimates by  $\hat{\beta}^{(r)}$  and using (5) we obtain the model

$$y_t = x'_t \beta + \phi \hat{u}_{t-1}^{(r)} + u_t, \quad t = 2, 3, \dots, T, \quad (7)$$

where

$$\hat{u}_t^{(r)} = y_t - x'_t \hat{\beta}^{(r)}, \quad t = 2, 3, \dots, T$$

are the residuals computed from the restricted model. We find that the one step Gauss-Newton estimate of the auto-regressive parameter  $\phi$  appears as the regression parameter for the constructed variable  $\hat{u}_{t-1}^{(r)}$ , appearing in the linear model (7). The  $L_1$ -norm criterion function to be minimized is therefore

$$S_0(\beta, \phi) = \sum_{t=2}^T |y_t - x'_t \beta - \hat{u}_{t-1}^{(r)} \phi|.$$

There are at least three immediate implications. First, the significance of including the auto-regressive error process to the model can be analyzed graphically by an added variable plot. Secondly, the significance of the auto-regressive process can formally be tested using the Lagrange multiplier test, by testing the significance of the regression parameter  $\phi$  of the linear regression model (7). Thirdly, the model (7) can be estimated and the gradient for the Lagrange multiplier test can be computed by using standard routines designed for linear models.

The asymptotic distribution of  $L_1$ -norm estimators of parameters in linear regression models was derived by Bassett and Koenker (1978) and Amemiya (1982). An application of this theory to the  $L_1$ -norm estimator  $(\hat{\beta}, \hat{\phi})'$  of the parameters of the linear model (7) yields that  $\sqrt{T}(\hat{\beta} - \beta', \hat{\phi})'$  is asymptotically normally distributed with mean zero and covariance matrix

$$\omega^2 \begin{pmatrix} D^{-1} & 0 \\ 0 & d^{-1} \end{pmatrix},$$

where  $\omega^2 = 1/(2f(0))^2$ , provided  $H_0$  is true.

Following Koenker and Bassett (1982) we define

$$W(\delta_1, \delta_2) = \sum_{t=2}^T \left| u_t - (x'_t \delta_1 + \hat{u}_{t-1}^{(r)} \delta_2) / \sqrt{T-1} \right|,$$

so that  $W(\sqrt{T-1}(\hat{\beta}^{(r)} - \beta), \sqrt{T-1}(\hat{\phi} - \phi)) = S_0(\hat{\beta}^{(r)}, \hat{\phi})$ . We further define the normalized gradient of  $W$  as

$$g(\delta_1, \delta_2) = \frac{1}{\sqrt{T-1}} \sum_{t=2}^T \begin{pmatrix} x_t \\ \hat{u}_{t-1}^{(r)} \end{pmatrix} \text{sign} \left( u_t - (x'_t \delta_1 + \hat{u}_{t-1}^{(r)} \delta_2) / \sqrt{T-1} \right).$$

Evaluating  $g$  at  $(\hat{\delta}_1, \hat{\delta}_2) = (\sqrt{T-1}(\hat{\beta}^{(r)} - \beta), -\sqrt{T-1}\phi)$  yields the vector  $g(\hat{\delta}_1, \hat{\delta}_2) = (\hat{g}_1, \hat{g}_2)'$ , where

$$\hat{g}_2 = \frac{1}{\sqrt{T-1}} \sum_{t=2}^T \hat{u}_{t-1}^{(r)} \text{sign}(\hat{u}_t^{(r)}).$$

If  $\hat{g}_2$  is large then  $H_0$  is implausible. The Lagrange multiplier test statistic is now defined as the quadratic form of the gradient

$$\xi = \frac{\hat{g}_2^2}{d} = \frac{\left\{ \sum_{t=2}^T \hat{u}_{t-1}^{(r)} \text{sign}(\hat{u}_t^{(r)}) \right\}^2}{(T-1)d}$$

From Koenker and Bassett (1982) it follows that the distribution of  $\xi$  is approximately non-central  $\chi^2$  with one degree of freedom and non-centrality parameter  $(T-1)\frac{d}{\omega^2}\phi^2$ . In particular, when  $H_0$  is true the asymptotic distribution of  $\xi$  is the central  $\chi^2$  distribution with one degree of freedom.

A comparison of the powers of the  $L_1$ -norm based Lagrange multiplier test and the corresponding test based on least squares can be done in terms of asymptotic relative efficiency. This quantity is the ratio of the noncentrality parameters of the limiting distributions,  $ARE = d/\omega^2$ . This comparison is therefore similar to a comparison of estimators efficiency. For normally distributed errors  $ARE = 2/\pi \approx 0.64$  so that the least squares method is preferable. However, as the error term variance increases and  $\omega^2$  decreases, the  $L_1$ -norm based methods become more favorable in terms of  $ARE$ . At the Laplace distribution, for example,  $ARE = 2$ .

## 4 Final remarks

This paper describes a Lagrange multiplier test for testing for a first-order auto-regressive error process in a linear regression model. There are several immediate extensions of the procedure. First, since the variance  $d$  of the disturbances  $v_t$  is unknown in most of the applications, it needs to be replaced by an estimator in the definition of  $\xi$ . A possible estimator is obtained by first estimating  $\phi$  by minimizing  $S_0$  and then defining

$$\hat{d} = \frac{1}{T-1} \sum_{t=2}^T \left( \hat{u}_t^{(r)} - \hat{\phi} \hat{u}_{t-1}^{(r)} \right)^2.$$

Secondly, Assumption 1 can be weakened to

$$\text{Assumption 1': } \max_{1 \leq t \leq T} x_t' (X'X)^{-1} x_t \rightarrow 0, \text{ as } T \rightarrow \infty.$$

This would change the normalization of  $\hat{\beta} - \beta$  from  $\sqrt{T-1}$  to  $(X'X)^{-1/2}$ . Thirdly, a test for a  $p$ -th order auto-regressive error process

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \dots + \phi_p u_{t-p} + v_t, \quad t = 1, 2, \dots, T,$$

is obtained if (1) is lagged by 1, 2, ...,  $p$  periods, multiplying the equations by  $\phi_1, \phi_2, \dots, \phi_p$ , respectively, and subtracting from (1) to obtain

$$\begin{aligned} & y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \dots - \phi_p y_{t-p} \\ = & \left( x_t' - \phi_1 x_{t-1}' - \phi_2 x_{t-2}' - \dots - \phi_p x_{t-p}' \right) \beta + v_t, \quad t = p+1, p+2, \dots, T. \end{aligned}$$

The  $L_1$ -norm criterion function, given the first  $p$  observations, is now

$$S(\beta, \phi) = \sum_{t=p+1}^T |y_t - h_t(\beta, \phi)|$$

with

$$h_t(\beta, \phi) =$$

$$\phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \left( x'_t - \phi_1 x'_{t-1} - \phi_2 x'_{t-2} - \dots - \phi_p x'_{t-p} \right) \beta.$$

The restricted model under the hypothesis that  $\phi_1 = \phi_2 = \dots = \phi_p = 0$  becomes

$$y_t = x'_t \beta + \phi_1 \hat{u}_{t-1} + \phi_2 \hat{u}_{t-2} + \dots + \phi_p \hat{u}_{t-p} + u_t, \quad t = p+1, p+2, \dots, T$$

and the Lagrange multiplier test of the hypothesis of no auto-regressive errors of order less than or equal to  $p$  is equivalent to the Lagrange multiplier test of the hypothesis that the regression coefficients  $\phi_1, \phi_2, \dots, \phi_p$  in the linear regression model are equal to zero.

Finally, it is worth pointing out that the approach for testing for auto-correlated error terms may be extended, under fairly mild regularity conditions, to the more general case of M-estimation of linear models.

## References

- [1] Amemiya, T. (1982). Two stage least absolute deviations estimators. *Econometrica* **50**, 689-712.
- [2] Anderson, D.H. and Osborne, M.R. (1977). Discrete, nonlinear approximation problems in polyhedral norms. *Numer. Math* **28**, 143-156.
- [3] Armstrong, R.D., Frome, E.L., and Kung, D.S. (1979). A revised simplex algorithm for the absolute deviation curve fitting problem. *Commun. Statist - Simul. Comput.* **8**, 175-190.
- [4] Barrodale, I. and Roberts, F.D.K. (1974). Algorithm 478: Solution of an overdetermined system of equations in the  $\ell_1$ -norm. *Commun. Assoc. Comput. Math.* **17**, 319-320.
- [5] Bassett, G. and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *J. Am. Statist. Assoc.* **73**, 618-622.
- [6] Cochrane, D. and Orcutt, G.H. (1949). Applications of least squares regressions to relationships concerning autocorrelated error terms. *J. Am. Statist. Assoc.* **44**, 32-61.
- [7] Coursey, D. and Nyquist, H. (1983) On least absolute error estimation of linear regression models with dependent stable residuals. *Rev. Econ. Statist.* **65**, 687-692.

- [8] Coursey, D. and Nyquist, H. (1986). A procedure for obtaining M-estimates in regression models with serially dependent errors. *J. Time Ser. Analys.* **7**, 255-267.
- [9] Koenker, R. and Bassett, G. (1982). Tests of linear hypotheses and  $\ell_1$  estimation. *Econometrica* **50**, 1577-1583.
- [10] Nyquist, H. (1992).  $L_1$ -norm estimation of regression models with serially dependent error terms. In  *$L_1$  Statistical Analysis and Related Methods*, Ed. Y. Dodge, pp. 253-264. Amsterdam: North-Holland.
- [11] Osborne, M.R. and Watson, G.A. (1971). On an algorithm for discrete nonlinear  $L_1$  approximation. *Comput. J.* **14**, 184-188.

# Gaussianity and nonlinearity of foreign exchange rates

**Nobuhiko Terui**

*Tohoku University, Kawauchi, Sendai, Japan*

**Takeaki Kariya**

*Hitotsubashi University, Kunitachi, Tokyo, Japan*

*Abstract:* In this paper, we test on the Gaussianity and nonlinearity of the foreign exchange rate return series by the Gaussianity test due to Kariya, Tsay, Terui and Li (1994) and by the five well-known nonlinearity tests for stationary time series. The daily returns of the foreign exchange rate we consider exhibit the strong non-Gaussianity or nonlinearity, but a central limit effect is observed with observational frequency longer even under stringent test.

*Key words:* KTTL test for Gaussianity, exchange rate, nonlinearity tests.

AMS subject classification: Primary 62M15; secondary 62M10.

## 1 Introduction

In economic time series analysis, Gaussianity is often assumed for modeling and for constructing asymptotic tests such as unit root test, cointegration test, e.t.c. or tabulating their critical points. In particular, in finance it is quite common to develop models and theories under the assumption of Brownianity or Gaussianity and apply them to real data. For example, the Black-Scholes stock option theory assumes that log prices follow a Brownian process or equivalently that returns of a stock follow a Gaussian process and the so-called CAPM (capital assets pricing market) model assumes normality for returns at least in their original forms. Although these theories have been developed in a less restrictive way for normality, most of empirical modelling and many time series tests frequently used for financial series

are constructed under Gaussianity. Therefore it is quite important to check whether our financial data is consistent with the assumption.

In this paper, applying the Gaussianity test proposed by Kariya, Tsay, Terui and Li (1994), shortened as the KTTL test<sup>1</sup> below, we test on the Gaussianity of the five foreign exchange rate returns and by considering the close relationship between the Gaussianity and linearity, we apply some nonlinearity tests to these data sets. The KTTL test of Gaussianity is a test that checks the consistency of the moment structure of a series with that of a Gaussian process up to an arbitrary order of the moments. In fact, the Gaussianity or equivalently normality is completely characterized by the moment structure (see Billingsley, 1986). It is also noted that the KTTL test is consistent with any stationary time series structure, and that such tests as the skewness test and the kurtosis test *e.t.c.* are not only partial tests, but also they assume *i.i.d.*-ness (independently-and-identically-distributed-ness). On the other hand, most of nonlinearity tests for stationary time series developed so far employ a linear process as a null hypothesis and set up, as alternatives, specific nonlinear models with additive noise. In fact, although there is a gap between non-Gaussianity and nonlinearity in stationary time series structure, induced time series from specific nonlinear model is almost always non-Gaussian as far as we assume the Gaussianity on the additive noise of nonlinear model, of which assumption is common in practical testing procedures. In this sense, nonlinearity tests constitute some kinds of the Gaussianity test.

We describe the KTTL test in Section 2. In Section 3, some nonlinearity testing methods for stationary time series are explained. Our empirical observations on the Gaussianity and nonlinearity of the return of foreign exchange rate are provided in Section 4. We observe the followings:

The daily series of the foreign exchange rate return show strong non-Gaussianity by the KTTL tests and nonlinearity by the five nonlinearity tests. The analysis of different observational intervals show the operation of a central limit effect in the sense that the *p*-values get larger as the observational period become longer for most of series. The KTTL omnibus test shows that the 6th order moment structures of monthly series are not consistent with those of a Gaussian process, although the moment structures up to the 4th order are not always incompatible with those of Gaussianity, where most of previous tests, including Jack-Bera test, etc., take only up to the 4th order moments into considerations.

---

<sup>1</sup>FORTRAN code for the KTTL test is available on request via e-mail : terui@econ.tohoku.ac.jp.

## 2 The KTTL test for Gaussianity

In this section, we describe the KTTL test of Gaussianity. To do so, we first summarize the KTTL test of multinormality, from which the KTTL test of Gaussianity follows.

Let  $\mathbf{Y} = (y_1, y_2, \dots, y_n)'$  follow multivariate normal distribution with mean vector  $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)'$  and covariance matrix  $\Sigma = (\sigma_{ij})$ , and denote the standardized variate of  $y_i$  as  $z_i = (y_i - \mu_i)/\sqrt{\sigma_{ii}}$ . Define  $w_j^{(p)} = h^{(p)}(z_j)$  as the  $p$ -th order Hermite polynomial of  $z_j$  for  $p = 1, \dots, P$  and  $j = 1, \dots, n$ . For example,  $w_j^{(1)} = z_j$ ,  $w_j^{(2)} = (z_j^2 - 1)/\sqrt{2}$ ,  $w_j^{(3)} = (z_j^3 - 3z_j)/\sqrt{6}$ ,  $w_j^{(4)} = (z_j^4 - 6z_j^2 + 3)/\sqrt{24}$  and so on.

Let

$$\psi_{ij}^{(p,q)} = \text{Cov}(w_i^{(p)}, w_j^{(q)}). \quad (1)$$

Then, under the null hypothesis of multinormality of  $\mathbf{Y}$ , it holds that for every  $i, j = 1, \dots, n$  and  $p, q = 1, \dots, P$ ,  $E(w_i^{(p)}) = 0$  and  $\psi_{ij}^{(pq)} = \phi_{ij}^{(pq)}$  with

$$\phi_{ij}^{(p,q)} = \begin{cases} \phi_{ij}^p & p = q \\ 0 & p \neq q, \end{cases} \quad (2)$$

where  $\phi_{ij} = \text{Cov}(w_i^{(1)}, w_j^{(1)})$ . This characterization is shown in Kendall and Stuart (1964, p.600). Granger and Newbold (1986, p.308) discussed its application to time series as a technique of instantaneous transformation.

Let  $\mathbf{Y}_t, t = 1, \dots, T$ , be the vectors of observation of  $\mathbf{Y}$  and define the sample version of  $z_i$  as  $u_{it} = (y_{it} - \bar{y}_i)/\sqrt{s_{ii}}$ , where  $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$  and  $s_{ij} = \frac{1}{T} \sum_{t=1}^T (y_{it} - \bar{y}_i)(y_{jt} - \bar{y}_j)$ . And denote the corresponding transformed variates of  $u_{it}$  via Hermite polynomials as  $v_{it}^{(p)} = h^{(p)}(u_{it})$  for  $i = 1, \dots, n$  and  $p = 1, \dots, P$ . Under the assumption of the existence of the  $2P$ th order moments of  $y_1, y_2, \dots, y_n$ , where  $P$  is the maximum order of the Hermite transformation under consideration and fixed in advance, two kinds of estimator for the covariance between  $w_i^{(p)}$  and  $w_j^{(q)}$  are used to construct an asymptotic test of multinormality. One is a consistent estimator only under the null hypothesis of multinormality, and the other is a consistent estimator under any arbitrary distribution. The former is given by

$$\hat{\phi}_{ij}^{(p,q)} = \begin{cases} (\hat{\phi}_{ij})^p & p = q \\ 0 & p \neq q, \end{cases} \quad (3)$$

where  $\hat{\phi}_{ij}$  is the sample correlation coefficient between  $z_i = w_i^{(1)}$  and  $z_j = w_j^{(1)}$ , and the latter is the conventional sample covariance estimate of  $\psi_{ij}^{(p,q)}$ ;



$$r_{ij}^{(pq)} = \frac{1}{T} \sum_{t=1}^T v_{it}^{(p)} v_{jt}^{(q)}. \quad (4)$$

The proposed test employs the difference between (3) and (4) as a test statistic.

Define a  $Pn \times Pn$  symmetric matrix  $\mathfrak{R}$  by

$$\mathfrak{R} = \begin{bmatrix} \mathbf{R}^{(1,1)} & \mathbf{R}^{(1,2)} & \cdots & \mathbf{R}^{(1,P)} \\ \mathbf{R}^{(2,1)} & \mathbf{R}^{(2,2)} & \cdots & \mathbf{R}^{(2,P)} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \mathbf{R}^{(P,1)} & \mathbf{R}^{(P,2)} & \cdots & \mathbf{R}^{(P,P)} \end{bmatrix}, \quad (5)$$

where the  $\mathbf{R}^{(p,q)}$  is an  $n \times n$  matrix whose  $(i, j)$  element is  $r_{ij}^{(p,q)}$ . Now compose  $Pn(Pn + 1)/2$  dimensional vector  $\mathbf{r}_P$  as follows:

$$\mathbf{r}^{(p,p)} = \text{Vech}(\mathbf{R}^{(p,p)}) : n(n + 1)/2 \times 1 \quad (6)$$

$$\mathbf{r}^{(p,q)} = \text{Vec}(\mathbf{R}^{(p,q)}) : n^2 \times 1 \quad (7)$$

$$\mathbf{r}_{P_1} = (\mathbf{r}^{(1,1)'}, \mathbf{r}^{(2,2)'}, \dots, \mathbf{r}^{(P,P)'})' : f_1 = Pn(n + 1)/2 \times 1 \quad (8)$$

$$\mathbf{r}_{P_2} = (\mathbf{r}^{(1,2)'}, \mathbf{r}^{(1,3)'}, \dots, \mathbf{r}^{(2,3)'}, \dots, \mathbf{r}^{(P-1,P)'})' : f_2 = P(P - 1)n^2/2 \times 1 \quad (9)$$

$$\mathbf{r}_P = (\mathbf{r}_{P_1}', \mathbf{r}_{P_2}')' : f = f_1 + f_2 = Pn(Pn + 1)/2 \times 1 \quad (10)$$

Here for any  $n \times n$  symmetric matrix  $\mathbf{A} = (a_{ij})$ ,

$$\text{Vech}(\mathbf{A}) = (a_{11}, a_{12}, \dots, a_{1n}; a_{22}, a_{23}, \dots, a_{2n}; \dots; a_{nn})$$

and for any  $n \times n$  matrix  $\mathbf{B} = (b_{ij})$ ,

$$\text{Vec}(\mathbf{B}) = (b_{11}, b_{12}, \dots, b_{1n}; b_{21}, b_{22}, \dots, b_{2n}; \dots; b_{n1}, b_{n2}, \dots, b_{nn}).$$

Correspondently, by replacing  $r_{ij}^{(p,q)}$  by  $\psi_{ij}^{(p,q)}$ , we can derive an  $f \times 1$  parameter vector  $\boldsymbol{\psi}_P$  under the alternative hypothesis. Also under the null hypothesis, in the same way we define, by replacing  $r_{ij}^{(p,q)}$  by  $\hat{\phi}_{ij}^{(p,q)}$  and  $\phi_{ij}^{(p,q)}$ , we can define

$$\hat{\boldsymbol{\phi}}_P = (\hat{\boldsymbol{\phi}}_{P_1}', \mathbf{O}')' \quad \text{and} \quad \boldsymbol{\phi}_P = (\boldsymbol{\phi}_{P_1}', \mathbf{O}')', \quad (11)$$

which correspond to  $\mathbf{r}_P$  and  $\boldsymbol{\psi}_P$  respectively. Then, assuming the  $2P$ th order moment of  $y_{it}$ ,  $\mathbf{r}_P$  follows the asymptotic multivariate normal distribution,

$$\sqrt{T}(\mathbf{r}_P - \boldsymbol{\psi}_P) \sim N_f(\mathbf{O}, \boldsymbol{\Lambda}) \quad (\boldsymbol{\Lambda} \text{ is the function of } \boldsymbol{\psi}_P). \quad (12)$$

Similarly it follows from (3) under the null hypothesis that

$$\sqrt{T}(\hat{\boldsymbol{\phi}}_P - \boldsymbol{\phi}_P) \sim N_f(\mathbf{O}, \boldsymbol{\Lambda}^*) \quad (\boldsymbol{\Lambda}^* \text{ is a function of } \phi_{ij}^{(1,1)}). \quad (13)$$

The KTTL test detects the equivalence between  $\boldsymbol{\psi}_P$  and  $\boldsymbol{\phi}_P$  under the null hypothesis based on the difference of  $\mathbf{r}_P$  and  $\hat{\boldsymbol{\phi}}_P = (\hat{\boldsymbol{\phi}}_{P1}', \mathbf{O}')$ . As we have  $\{n(n+1)/2\}$  equivalent relationships;  $\hat{\phi}_{ij}^{(1,1)} \equiv r_{ij}^{(1,1)}$ , the test is  $g = \{f - n(n+1)/2\}$  dimensional vector

$$\mathbf{C}_{Po} = \mathbf{C}_{Po}(\mathbf{r}_P) = (\mathbf{C}'_{P1}, \mathbf{C}'_{P2})', \quad (14)$$

where

$$\mathbf{C}_{P1} = (\mathbf{c}^{(2,2)'}, \dots, \mathbf{c}^{(P,P)'})' \quad (15)$$

$$\mathbf{C}_{P2} = (\mathbf{c}^{(1,2)'}, \dots, \mathbf{c}^{(1,P)'}; \dots; \mathbf{c}^{(P-1,P)'})' \quad \text{and} \quad (16)$$

$$\mathbf{c}^{(p,q)} = \begin{cases} \mathbf{r}^{(p,p)} - \hat{\boldsymbol{\phi}}^{(p,p)} & p = q \\ \mathbf{r}^{(p,q)}, & p \neq q. \end{cases} \quad (17)$$

In fact,  $\mathbf{C}_{Po}$  is expressed as

$$\mathbf{C}_{Po} = [\mathbf{O} \quad \mathbf{I}](\mathbf{r}_P - \hat{\boldsymbol{\phi}}_P), \quad (18)$$

where  $\mathbf{O}$  is the  $(f - n(n+1)n/2) \times 1$  zero matrix, and  $\mathbf{I}$  is the  $(f - n(n+1)n/2)$  dimensional identity matrix. Under the null hypothesis, where  $\boldsymbol{\psi}_P = \boldsymbol{\phi}_P$ , it holds from (12) and (13) that

$$\sqrt{T}\mathbf{C}_{Po} \sim N_g(\mathbf{O}, \mathbf{J}(\boldsymbol{\phi}_P)\boldsymbol{\Lambda}\mathbf{J}(\boldsymbol{\phi}_P)'), \quad (19)$$

where the  $(f - n(n+1)/2) \times f$  matrix  $\mathbf{J}(\boldsymbol{\phi}_P)$  is

$$\begin{aligned} \mathbf{J}(\boldsymbol{\phi}_P) &= \left. \frac{\partial \mathbf{C}_{Po}}{\partial \mathbf{r}_P} \right|_{\mathbf{r}_P = \boldsymbol{\phi}_P} \\ &= \left( \frac{\partial \mathbf{c}^{(2,2)}}{\partial \mathbf{r}_P}, \dots, \frac{\partial \mathbf{c}^{(P,P)}}{\partial \mathbf{r}_P}; \frac{\partial \mathbf{c}^{(1,2)}}{\partial \mathbf{r}_P}, \dots, \frac{\partial \mathbf{c}^{(P-1,P)}}{\partial \mathbf{r}_P} \right) \bigg|_{\mathbf{r}_P = \boldsymbol{\phi}_P} \end{aligned} \quad (20)$$

and the matrix differential is given by

$$\frac{\partial \mathbf{c}^{(p,q)}}{\partial \mathbf{r}_P} = \left( \frac{\partial c_{ij}^{(p,q)}}{\partial r_{kl}^{(a,b)}} \right). \quad (21)$$

The KTTL test statistic for multinormality is a Wald type chi-square test defined by

$$W_o = T \mathbf{C}'_{Po} [\hat{\mathbf{J}} \hat{\mathbf{\Lambda}} \hat{\mathbf{J}}']^{-1} \mathbf{C}_{Po}, \quad (22)$$

where  $\hat{\mathbf{J}} = J(\hat{\phi}_P)$  and  $\hat{\mathbf{\Lambda}} = (\hat{\lambda}_{ij,kl}^{(pq,ab)})$  with

$$\hat{\lambda}_{ij,kl}^{(pq,ab)} = \frac{1}{T} \sum_{i=1}^T (v_{it}^{(p)} v_{jt}^{(q)} - r_{ij}^{(pq)}) (v_{kt}^{(a)} v_{lt}^{(b)} - r_{kl}^{(ab)}). \quad (23)$$

The KTTL test of Gaussianity for univariate series is a modification of the above test. Let  $\{x_t\}$  be a univariate stationary process with  $E(x_t) = \mu$  and  $\text{Cov}(x_t, x_{t-k}) = \gamma_k$  and assume the mixing condition  $\sum_{k=-\infty}^{\infty} |k| |\gamma_k| < \infty$ . Then the following two methods for constructing a test are proposed.

[I] Overlapping method

Set, for  $i = 1, \dots, n$ ,

$$y_{it} = x_{t-i+1}, \quad (24)$$

and corresponding to *i.i.d.* case, we define  $z_{it} = (y_t - \mu)/\sqrt{\gamma_0} = (x_{t-i+1} - \mu)/\sqrt{\gamma_0}$ .

[II] Non-overlapping method

For a given positive integer  $n$  and the realizations  $\{x_1, \dots, x_N\}$ , set  $n$ -dimensional non-overlapping random vectors  $\{y_t = (y_{1t}, \dots, y_{nt})'\}$  with

$$y_{it} = x_{n(t-1)+i}, i = 1, \dots, n; t = 1, \dots, T, \quad (25)$$

where  $T = [N/n]$  is the integer part of  $N/n$ . Then we define  $z_{it} = (x_{n(t-1)+i} - \mu)/\sqrt{\gamma_0}$ .

For both [I] and [II], it follows that  $\text{Cov}(z_{it}, z_{jt}) = \gamma_{i-j}/\gamma_0 = \rho_{i-j} = \phi_{ij}$ . Defining  $w_{it}^{(p)} = h^{(p)}(z_{it})$ , just like *i.i.d.* case, under the Gaussianity of  $x_t$ , it holds that  $\psi_{ij}^{(p,q)} = \phi_{ij}^{(p,q)}$  with

$$\phi_{ij}^{(pq)} = \begin{cases} (\phi_{ij})^p & p = q \\ 0 & p \neq q. \end{cases} \quad (26)$$

The sample variates of  $z_{it}$  and  $w_{it}^{(p)}$  are defined as  $u_{it} = (y_{it} - \bar{y}_i)/\sqrt{s_{ii}}$ ,  $v_{it}^{(p)} = h^{(p)}(u_{it})$  for  $i = 1, \dots, n$  and  $p = 1, \dots, P$ , where  $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$  and

$s_{ij} = \frac{1}{T} \sum_{t=\max(i,j)}^T (y_{it} - \bar{y}_i)(y_{jt} - \bar{y}_j)$ . For time series data, the estimates corresponding to (3) and (4) should be respectively modified as

$$r_{ij}^{(pq)} = \frac{1}{T} \sum_{t=\max(i,j)}^T v_{it}^{(p)} v_{jt}^{(q)} \quad (27)$$

and

$$\hat{\phi}_{ij}^{(pq)} = \begin{cases} \hat{\rho}_{i-j}^p & p = q \\ 0 & p \neq q \end{cases} \quad (28)$$

for  $i, j = 1, \dots, n : p, q = 1, \dots, P$ . Under the Gaussianity of  $\{x_t\}$  with the mixing condition above, Keenan (1983) proved the asymptotic normality that  $\sqrt{T}(\mathbf{r}_P - \boldsymbol{\phi}_P) \rightarrow N(\mathbf{0}, \mathbf{J}\boldsymbol{\Lambda}\mathbf{J}')$ . Therefore, exactly following the argument of the *i.i.d.* case, we obtain the asymptotically  $\chi^2$  test statistic (22) with d.f.  $\{f - n(n+1)/2\}$  for testing the Gaussianity of a stationary time series.

The test in (22) test is an omnibus test which detects departures from Gaussianity, and it can be decomposed into two parts. The first part tests departures from the even moment structure when  $p = q$  and the second part tests departures from the odd moment structure when  $p \neq q$ . It is useful to have a separate test for each part when we are interested in the symmetry and the tail behavior of the underlying distribution separately. Corresponding to the dimensions of  $\mathbf{C}_{P1}$  and  $\mathbf{C}_{P2}$ , the appropriate decomposition of  $\boldsymbol{\Lambda}$  and  $\mathbf{J}$  produces the following test statistics:

$$W_1 = T\mathbf{C}'_{P1} [\hat{\mathbf{J}}_{11} \hat{\boldsymbol{\Lambda}}_{11} \hat{\mathbf{J}}'_{11}]^{-1} \mathbf{C}_{P1}, \text{ and} \quad (29)$$

$$W_2 = T\mathbf{C}'_{P2} [\hat{\boldsymbol{\Lambda}}_{22}]^{-1} \mathbf{C}_{P2}. \quad (30)$$

Under the null hypothesis of Gaussianity, the asymptotic distributions of these two test statistics are  $\chi^2$  with degrees of freedom  $f_1 - n(n+1)/2$  and  $f_2$  respectively. In fact, setting  $\mathbf{A} \equiv [\hat{\mathbf{J}} \hat{\boldsymbol{\Lambda}} \hat{\mathbf{J}}']^{-1}$ , and decomposing  $\mathbf{A}$  appropriately according to the dimension of  $\mathbf{C}_{P1}$  and  $\mathbf{C}_{P2}$ , then we have the following relationship among  $W_o$ ,  $W_1$  and  $W_2$ ;

$$\begin{aligned} W_o &= T\mathbf{C}'_{Po} [\hat{\mathbf{J}} \hat{\boldsymbol{\Lambda}} \hat{\mathbf{J}}']^{-1} \mathbf{C}_{Po} \\ &\equiv T(\mathbf{C}_{P1}', \mathbf{C}'_{P2}) \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12} & \mathbf{A}_{22} \end{pmatrix} (\mathbf{C}_{P1}, \mathbf{C}_{P2}) \\ &= T\mathbf{C}'_{P1} \mathbf{A}_{11} \mathbf{C}_{P1} + T\mathbf{C}'_{P2} \mathbf{A}_{22} \mathbf{C}_{P2} + 2T\mathbf{C}'_{P1} \mathbf{A}_{12} \mathbf{C}_{P2} \\ &= W_1 + W_2 + 2\mathbf{C}_{12}, \end{aligned} \quad (31)$$

where  $\mathbf{C}_{12} = T\mathbf{C}'_{P1}\mathbf{A}_{12}\mathbf{C}_{P2}$ . Now considering here  $\frac{\mathbf{C}'_{P1}\mathbf{A}_{12}\mathbf{C}_{P2}}{\sqrt{W_1}\sqrt{W_2}}$  may be regarded as a kind of correlation between  $\mathbf{C}_{P1}$  and  $\mathbf{C}_{P2}$ . In fact, in the case where both  $\mathbf{C}_{P1}$  and  $\mathbf{C}_{P2}$  are one dimensional vector,  $C_{12}/(\sqrt{W_1}\sqrt{W_2})$  is shown to be equivalent to negative correlation between  $C_1$  and  $C_2$ .

### 3 Some nonlinearity tests

Let

$$X_t = h(X_{t-1}, X_{t-2}, \dots, X_{t-p}) + e_t \quad (32)$$

be an autoregressive nonlinear time series model, where  $\{e_t\}$  is *i.i.d.* with mean zero. If we assume the innovation  $e_t$  as Gaussian, nonlinearity test is equivalent to Gaussianity test. Here we use the five well known nonlinearity tests:

- (i) Ori-F test by Tsay (1986)
- (ii) Aug-F test by Luukkonenn, Saikkonen and Teräsvirta (1988)
- (iii) CUSUM test by Petruccielli and Davis (1986)
- (iv) TAR-F test by Tsay (1989)
- (v) New-F test by Tsay (1988)

All of these tests set up, as a null hypothesis, a linear process. Based on the Volterra expansion of (32) around  $\mathbf{O} = (0, 0, \dots)'$

$$x_t = \mu + \sum_{u=1}^{\infty} \psi_u x_{t-u} + \sum_{u,v=1}^{\infty} \psi_{uv} x_{t-u} x_{t-v} + \sum_{u,v,w=1}^{\infty} \psi_{uvw} x_{t-u} x_{t-v} x_{t-w} + \dots + e_t, \quad (33)$$

where

$$\begin{aligned} \mu &= h(\mathbf{O}), \phi_u = \left. \frac{\partial h}{\partial x_{t-u}} \right|_{\mathbf{O}}, \phi_{uv} = \left. \frac{\partial^2 h}{\partial x_{t-u} \partial x_{t-v}} \right|_{\mathbf{O}}, \phi_{uvw} \\ &= \left. \frac{\partial^3 h}{\partial x_{t-u} \partial x_{t-v} \partial x_{t-w}} \right|_{\mathbf{O}}, \text{e.t.c.}, \end{aligned}$$

the Ori-F and Aug-F tests detect against the nonlinearity of the second and third order polynomials respectively. The CUSUM, TAR-F and New-F tests assume the threshold type nonlinear alternatives;

$$x_t = \beta_0^{(j)} + \sum_{i=1}^p \beta_i^{(j)} x_{t-i} + a_t^{(j)} \quad (j = 1, 2), \quad (34)$$

where  $\{a_t^{(j)}\}$  is the innovation of mean zero and variance  $\sigma_j^2$ . The New-F test covers the most extensive alternatives of nonlinearity, including ExpAR

proposed by Haggan and Ozaki (1981)

$$X_t = \{\phi_1 + \pi_1 \exp(-\gamma X_{t-1}^2)\}X_{t-1} + \{\phi_2 + \pi_2 \exp(-\gamma X_{t-1}^2)\}X_{t-2} + \cdots + \{\phi_p + \pi_p \exp(-\gamma X_{t-1}^2)\}X_{t-p} + e_t \quad (35)$$

and bilinear models by Granger and Anderson (1978), Subba Rao and Gabr (1984) and others

$$X_t + \sum_{j=1}^p a_j X_{t-j} = \sum_{j=0}^r c_j e_{t-j} + \sum_{i=1}^m \sum_{j=1}^k b_{ij} X_{t-i} e_{t-j}. \quad (36)$$

The detailed procedures and distributional properties regarding these tests are found in Granger and Teräsvirta (1993).

In order to implement these tests, the order  $p$  of autoregression for all the tests and the value of delay parameter  $d$  for the tests (iii), (iv), and (v) need to be specified. We set the maximum of  $p$  as 10 and let  $d$  run from 1 to 10. Each of nonlinearity tests with different set of  $(p, d)$  brings out different results. We employ the most significant result of the test among all the combinations of  $(p, d)$ .

## 4 Foreign exchange rate

Since the work of Westerfield (1977), a battery of research about the foreign exchange rate has been done. For a survey, see Levich (1985), Isard (1988), Mills (1993) and Campbell, Lo and McKilay (1997). Many of those works, for example, So (1987), Wolff (1987), Enders (1988) and etc., employ the Gaussianity of the process as a basic assumption. The careful investigations of this assumption should be needed in advance.

In this section, we investigate the Gaussianity and nonlinearity of five foreign exchange rate return series against US dollar: FRF (French Franc), JPY (Japanese Yen), CHF (Swiss Franc), GBR (English Starling), DEM (German Mark). We deal with three kinds of data; daily (1992.1.1.- 1993.12.31. (552 samples)), weekly (1984.1.2. - 1993.12.27. (521 samples: Mondays)) and monthly (1983.10.31. - 1994.10.31. (132 samples)). Over the observational period, we assume a homogeneity between these data.

We define a exchange rate return  $x_t$  at time  $t$  as  $x_t = \log R_t - \log R_{t-1}$  where  $R_t$  is a exchange rate at  $t$ .

### 4.1 Gaussianity

In this article, we use the overlapping method of the KTTL test because macroeconomic time series usually do not have enough numbers of the sample. Denote the KTTL test with  $P = 2$  and  $n = 1$  by O-21 and so on.

Applying the KTTL test to the data, we move  $P = 2, 3$  and  $n = 1, \dots, 4$  and the p-values of the tests O-21 through O-32 are tabulated in Table 1. The other tests with greater than  $P = 3$  and  $n = 2$  (that is, O-33 and O-34) reject the Gaussianity very strongly. In the following, we enumerate the empirical findings from the results.

### 1) Daily series

a) All the tests applied to other than CHF series reject the Gaussianity. The marginal moment test  $W_1$  rejects the Gaussianity with 5% significant level for all the series. In fact, the maximum value of the p-values is 0.03955 for CHF.

b) In O-21 and O-22 tests, the results of the  $W_2$  test are not significant for many cases.

c) The  $W_0$  test can not reject the Gaussianity of CHF even when  $P$  and  $n$  are large.

### 2) Weekly series

a) The Gaussianity of JPY is rejected more strongly than other series.

b) Compared with daily series, the number of significant series increases.

### 3) Monthly series

a) In case of  $P = 2$ , not only the omnibus test  $W_0$ , but also the marginal tests  $W_1$  and  $W_2$  can not reject the Gaussianity for all the series.

b) Compared with daily and weekly series, many p-values of all the tests,  $W_0$ ,  $W_1$  and  $W_2$ , are larger.

c) The  $W_1$  and  $W_2$  tests reject the Gaussianity of GBR strongly, however the  $W_0$  test can not necessarily be inconsistent with the Gaussianity (O-22, O-23 and O-31).

d) The  $W_0$  test rejects the Gaussianity more strongly when we set  $P = 3$ . This means that the 6th order moment structures of data sets are not consistent with those of a Gaussian process, although the moment structures up to the 4th order are not always incompatible with those of Gaussianity.

e) In case of  $P = 3$ , we have some cases where the omnibus test  $W_0$  reject the null hypothesis more strongly than the marginal tests  $W_1$  and  $W_2$  (JPY, CHF, DEM of O-31 and JPY, CHF of O-32). This means that  $2C_{12}$  of (21) takes the large positive values for those cases.

As a whole, we observe that, given  $P$ , the hypothesis of Gaussianity tends to be rejected more strongly, as  $n$  gets larger and in case of  $P = 3$ , and the Gaussianity is completely rejected for almost returns, especially in case of  $n \geq 2$ . Further we observe that a central limit effect is working on weekly series in the sense that the p-values get larger as the observational

period become longer.

Table 1: The KTTL Tests: Foreign Exchange Rate Returns.

| Variable                                    | Daily(552 samples) |          |          | Weekly(521 samples) |          |          | Monthly(132 samples) |          |          |
|---------------------------------------------|--------------------|----------|----------|---------------------|----------|----------|----------------------|----------|----------|
|                                             | $P(W_o)$           | $P(W_1)$ | $P(W_2)$ | $P(W_o)$            | $P(W_1)$ | $P(W_2)$ | $P(W_o)$             | $P(W_1)$ | $P(W_2)$ |
| O-21(D.F.: $W_o : 2, W_1 : 1, W_2 : 1$ )    |                    |          |          |                     |          |          |                      |          |          |
| FRF                                         | 0.01311            | 0.00622  | 0.57667  | 0.11414             | 0.02135  | 0.35430  | 0.47818              | 0.28592  | 0.09538  |
| JPY                                         | 0.01405            | 0.00244  | 0.35201  | 0.00674             | 0.00112  | 0.05900  | 0.66199              | 0.69024  | 0.49823  |
| CHF                                         | 0.14530            | 0.03956  | 0.84395  | 0.42304             | 0.23533  | 0.72896  | 0.72221              | 0.79985  | 0.37821  |
| GBR                                         | 0.00355            | 0.00006  | 0.15180  | 0.26114             | 0.00262  | 0.03956  | 0.26506              | 0.11417  | 0.77253  |
| DEM                                         | 0.01549            | 0.00567  | 0.73782  | 0.22792             | 0.07984  | 0.74623  | 0.55148              | 0.53422  | 0.16216  |
| O-22(D.F.: $W_o : 7, W_1 : 3, W_2 : 4$ )    |                    |          |          |                     |          |          |                      |          |          |
| FRF                                         | 0.00076            | 0.00020  | 0.11010  | 0.05483             | 0.00034  | 0.45788  | 0.75493              | 0.41573  | 0.06835  |
| JPY                                         | 0.01096            | 0.00000  | 0.01128  | 0.00305             | 0.00000  | 0.06881  | 0.87798              | 0.88747  | 0.81214  |
| CHF                                         | 0.20441            | 0.01291  | 0.36325  | 0.63944             | 0.16946  | 0.43741  | 0.94473              | 0.75097  | 0.73328  |
| GBR                                         | 0.00083            | 0.00000  | 0.00000  | 0.35538             | 0.00002  | 0.02632  | 0.53963              | 0.00000  | 0.00006  |
| DEM                                         | 0.02160            | 0.00128  | 0.81563  | 0.07231             | 0.00989  | 0.96215  | 0.78226              | 0.78435  | 0.17848  |
| O-23(D.F.: $W_o : 15, W_1 : 6, W_2 : 9$ )   |                    |          |          |                     |          |          |                      |          |          |
| FRF                                         | 0.00037            | 0.00000  | 0.00090  | 0.02444             | 0.00000  | 0.29071  | 0.84104              | 0.22524  | 0.00350  |
| JPY                                         | 0.01999            | 0.00000  | 0.00439  | 0.00126             | 0.00000  | 0.08059  | 0.85650              | 0.25579  | 0.83568  |
| CHF                                         | 0.28879            | 0.00219  | 0.09646  | 0.71675             | 0.16591  | 0.52572  | 0.98131              | 0.78192  | 0.81381  |
| GBR                                         | 0.00019            | 0.00000  | 0.00000  | 0.47406             | 0.00000  | 0.00168  | 0.73477              | 0.00000  | 0.00000  |
| DEM                                         | 0.03287            | 0.00014  | 0.69558  | 0.04213             | 0.00118  | 0.99706  | 0.81120              | 0.89435  | 0.08016  |
| O-24(D.F.: $W_o : 26, W_1 : 10, W_2 : 16$ ) |                    |          |          |                     |          |          |                      |          |          |
| FRF                                         | 0.00026            | 0.00000  | 0.00000  | 0.00261             | 0.00000  | 0.22997  | 0.72710              | 0.07933  | 0.00000  |
| JPY                                         | 0.00819            | 0.00000  | 0.00005  | 0.00016             | 0.00000  | 0.01655  | 0.87774              | 0.09242  | 0.75604  |
| CHF                                         | 0.43501            | 0.00029  | 0.03210  | 0.48250             | 0.01646  | 0.60762  | 0.87173              | 0.83366  | 0.48292  |
| GBR                                         | 0.00002            | 0.00000  | 0.00000  | 0.48250             | 0.01646  | 0.60762  | 0.17422              | 0.00000  | 0.00000  |
| DEM                                         | 0.02310            | 0.00000  | 0.20145  | 0.01875             | 0.00006  | 0.99967  | 0.67201              | 0.87080  | 0.01459  |
| O-31(D.F.: $W_o : 5, W_1 : 2, W_2 : 3$ )    |                    |          |          |                     |          |          |                      |          |          |
| FRF                                         | 0.00001            | 0.00000  | 0.00000  | 0.30700             | 0.00001  | 0.00636  | 0.0 0000             | 0.33776  | 0.00000  |
| JPY                                         | 0.00000            | 0.00000  | 0.00000  | 0.00002             | 0.00000  | 0.00000  | 0.00 012             | 0.07334  | 0.03814  |
| CHF                                         | 0.09266            | 0.00000  | 0.00000  | 0.86862             | 0.04989  | 0.34839  | 0.0 6818             | 0.18915  | 0.38700  |
| GBR                                         | 0.00004            | 0.00000  | 0.00000  | 0.00573             | 0.00000  | 0.00000  | 0.2 5814             | 0.00000  | 0.00001  |
| DEM                                         | 0.00973            | 0.00000  | 0.00023  | 0.35201             | 0.00266  | 0.31204  | 0.0 0000             | 0.61632  | 0.00000  |
| O-32(D.F.: $W_o : 18, W_1 : 6, W_2 : 12$ )  |                    |          |          |                     |          |          |                      |          |          |
| FRF                                         | 0.00000            | 0.00000  | 0.00000  | 0.22459             | 0.00000  | 0.00239  | 0.00000              | 0.01589  | 0.00000  |
| JPY                                         | 0.00000            | 0.00000  | 0.00000  | 0.00000             | 0.00000  | 0.00000  | 0.00000              | 0.00266  | 0.00000  |
| CHF                                         | 0.13266            | 0.00000  | 0.00000  | 0.45164             | 0.00005  | 0.03728  | 0.00217              | 0.15251  | 0.13535  |
| GBR                                         | 0.00000            | 0.00000  | 0.00000  | 0.00009             | 0.00000  | 0.00000  | 0.00000              | 0.00000  | 0.00000  |
| DEM                                         | 0.00429            | 0.00000  | 0.00000  | 0.51822             | 0.00008  | 0.52567  | 0.00000              | 0.18092  | 0.00000  |

$P(W_i)$  means the p-value of the test  $W_i, i = 0, 1, 2$ .

## 4.2 Nonlinearity

Table 2 shows the results of applying the five nonlinearity tests explained in Section 3 to the foreign exchange returns. The CUSUM test produces very different results from other tests. In Terui and Kariya (1996), it is shown that, applying these tests to 214 Japanese stock returns, the empirical distributions of the p-values of the CUSUM test have a very different shape of distribution function from those of four other tests. Further the simulation studies by Tsay (1988, 1989) indicate the low power of the CUSUM test and we can see this observation as an empirical evidence of it. Therefore we leave out the results of the CUSUM test.

From the table, we observe;

### 1) Daily series

a) FRF and GBR are non-Gaussian judging from all the nonlinearity tests.



b) JPY is incompatible with a Gaussian process for all the tests. It is possible for JPY to have a nonlinearity other than what the five tests assume as alternative, because the KTTL test also supports the non Gaussianity.

c) For CHF, the Ori-F and Aug-F tests support the Gaussianity and the TAR-F and New-F tests suggests the non-Gaussianity. Therefore CHF can have a threshold type nonlinearity.

d) DEM is compatible with a Gaussian process by the Aug-F test and inconsistent by the TAR-F and New-F tests. The threshold type nonlinearity might exit for DEM.

## 2) Weekly series

a) FRF, CHF and DEM could have threshold type nonlinearity because the TAR-F and New-F tests reject the null hypothesis of linearity.

b) The non-Gaussianity is suggested for JPY by all the tests.

c) GBR is compatible with a Gaussian process for all the tests.

## 3) Monthly series

a) There are no returns with the p-values greater than 1%, except for the TAR-F test applied to JPY.

Table 2: Nonlinearity Tests: Foreign Exchange Rate Returns.

| Variable | Ori-F   | Aug-F   | TAR-F   | CUSUM   | New-F   |
|----------|---------|---------|---------|---------|---------|
| Daily    |         |         |         |         |         |
| FRF      | 0.00000 | 0.00000 | 0.00110 | 0.10151 | 0.00053 |
| JPY      | 0.14047 | 0.24706 | 0.25270 | 0.74493 | 0.05492 |
| CHF      | 0.46322 | 0.68311 | 0.02595 | 0.04235 | 0.02791 |
| GBR      | 0.00012 | 0.00003 | 0.00262 | 0.00920 | 0.00241 |
| DEM      | 0.03721 | 0.08136 | 0.00511 | 0.02198 | 0.02273 |
| Weekly   |         |         |         |         |         |
| FRF      | 0.08246 | 0.06449 | 0.00341 | 0.18164 | 0.05208 |
| JPY      | 0.01152 | 0.01331 | 0.00090 | 0.11705 | 0.00000 |
| CHF      | 0.46697 | 0.68680 | 0.02721 | 0.04223 | 0.01986 |
| GBR      | 0.06608 | 0.14518 | 0.11549 | 0.33748 | 0.07851 |
| DEM      | 0.06415 | 0.05298 | 0.00125 | 0.11915 | 0.00035 |
| Monthly  |         |         |         |         |         |
| FRF      | 0.07120 | 0.04798 | 0.11767 | 0.04712 | 0.10081 |
| JPY      | 0.04101 | 0.01300 | 0.00159 | 0.12351 | 0.03721 |
| CHF      | 0.03099 | 0.07967 | 0.10196 | 0.15927 | 0.04794 |
| GBR      | 0.17833 | 0.22842 | 0.04982 | 0.13293 | 0.02420 |
| DEM      | 0.10053 | 0.12918 | 0.15184 | 0.06969 | 0.11015 |

## References

- [1] Bilingsley, P.(1986). *Probability and Measure* (2nd ed). New York: Wiley.
- [2] Campbell, J., Lo, A. and C. McKilay (1997). *The Econometrics of Financial Markets*. Princeton University Press.
- [3] Chen, R., A.L. Monigomery, K. Xu, G.C.Tiao and R.S.Tsay(1990). *Nonlinear Time Series Analysis: A Software Package*. Graduate School of Business, University of Chicago.
- [4] Dickey, D.A. and W.A. Fuller(1981). Likelihood ratio statistics for autoregressive time series unit root. *Econometrica* **49**, 1057-1072.
- [5] Enders, W.(1988). ARIMA and cointegration tests of PPP under fixed and flexible exchange rate regimes. *Review of Economics and Statistics* **55**, 251-276.
- [6] Granger, C.W.J. and Anderson, A.P.(1978). *An Introduction to Bilinear Time Series Models*. Göttingen: Vandenhoeck and Ruprecht.
- [7] Granger, C.W.J. and P.Newbold(1986a). *Forecasting Economic Time Series*. Orland: Academic Press.
- [8] Granger, C.W.J. and P.Newbold(1986b). Forecasting transformed series. *J. R. Statist. Soc. B* **38**, 189-203.
- [9] Granger, C.W.J. and T.Teräsvirta(1993). *Modeling Nonlinear Economic Relationships*. New York: Oxford University Press.
- [10] Haggan,V. and T.Ozaki(1981). Modeling nonlinear random vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* **68**, 189-196.
- [11] Isard, P.(1988). Exchange rate modelling : an assessment to alternative approach. In *Empirical Macroeconomics for Interdependent Economies*, Eds. R.C. Bryent et al., Vol. 1, pp. 183-201. Brookings Institute, Washington D.C.
- [12] Kariya,T., R. Tsay, N. Terui and H. Li(1994). Tests for multinormality with applications of time series. Discussion Paper Series A No.264, The Institute of Economic Research, Hitotsubashi University.
- [13] Keenan,D.M.(1983). Limiting behavior of functionals of the sample spectral distribution. *Ann. Statist.* **11**,1206-1217.
- [14] Kendall, M.G. and A. Stuart(1964). *The Advanced Theory of Statistics*, Vol. 2. London: Griffin.
- [15] Levich, R.M.(1985). Empirical studies of foreign exchange rate: price behavior, rate determination and market efficiency. In *Handbook of International Economics*, Eds. R.W. Jones et al., Vol. 2. Amsterdam: North-Holland.
- [16] Luukkonenn, R., Saikkonen, P. and Teräsvirta, T.(1988). Testing lin-

- earity against smooth transition autoregressive models. *Biometrika*, **75**, 491-499.
- [17] Mills, T.C. (1993). *The Econometric Modelling of Financial Time Series*. Cambridge University Press.
  - [18] Petrucci, J. and Davis, N.(1986). A portmanteaus test for self-exciting threshold autoregressive-type nonlinearity in time series. *Biometrika* **73**, 687-694.
  - [19] Priestley, M.B.(1988). *Nonlinear and Non-Stationary Time Series Analysis*. London: Academic Press.
  - [20] So, J.C.(1987). The sub-Gaussian distribution of currency futures: stable Paretian or nonstationary ?. *Review of Economics and Statistics* **69**, 100-107.
  - [21] Subba Rao, T. and M.M.Gabr(1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Berlin: Springer-Verlag.
  - [22] Terui, N. and T.Kariya(1996). Testing Gaussianity and linearity of Japanese stock returnss. Working paper, Tohoku University.
  - [23] Terui, N. and T.Kariya(1997). Gaussianity and nonlinearity of foreign exchange rate and macroeconomic time series. Working paper TI97-004/4, Tinbergen Institute, Erasmus University.
  - [24] Tsay, R.S.(1986). Nonlinearity tests for time series. *Biometrika* **73**, 461-466.
  - [25] Tsay, R.S.(1988). Nonlinear time series analysis of Blowfly population. *Journal of Time Series Analysis* **9**, 246-263.
  - [26] Tsay, R.S.(1989). Testing and modeling threshold autoregressive process. *J. Am. Statist. Assoc.* **84**, 231-240.
  - [27] Westerfield, J.M.(1977). An examination of foreign exchange risk under fixed and floating rate regimes. *Journal of International Economics* **7**, 181-200.
  - [28] Wolff, C.C.P.(1987). Time-varying parameters and the out-of- sample forecasting performance of structural exchange models. *Journal of Business and Economic Statistics* **5**, 87-97.

# Rank plots in the affine invariant case

**B.M. Brown**

*University of Tasmania, Australia*

**T.P. Hettmansperger**

*Pennsylvania State University, USA*

**J. Möttönen and H. Oja**

*University of Oulu, Finland*

*Abstract:* The bivariate ranks and quantiles based on the bivariate affine equivariant median are considered. Correspondences between two different plots for bivariate data, the direct diagram and the Oja rank plot, are described. Several illustrative examples are given.

*Key words:* Affine invariance, affine equivariance, bivariate quantile, bivariate rank, multivariate median.

AMS subject classification: 62G30, 62F07, 62H99.

## 1 Introduction

Rank methods occupy a central role among standard univariate statistical methods, and form the backbone of conventional nonparametrics. Consequently, it has been recently of some interest to explore concepts of rank for multivariate data, and in particular, for bivariate data. There are various alternatives, including ideas based on depth (Liu, 1990,1992; Liu and Singh, 1993). But another analytic definition of bivariate rank which leads to appealing bivariate analogues of univariate rank statistical methods is reached through the gradients of a convex objective function used to define a bivariate median; see Brown and Hettmansperger (1987a,b), Hettmansperger, Nyblom and Oja (1992), Hettmansperger, Möttönen and Oja (1997a,b) and Möttönen and Oja (1995). To show how this idea works, the notion of

univariate rank is set up this way in Section 2.

There are several possible definitions of bivariate median which could be used to develop a notion of bivariate rank (Small, 1990; Niinimaa and Oja, 1997). Among these, the bivariate median of Oja (1983) is affine invariant. The resulting bivariate ranks are called Oja ranks, and defined in Section 3. They lead to the idea of bivariate quantile, which is a data item or region or chord between data items having prescribed rank.

The purpose of this paper is to examine the connections between and uses of two corresponding plots. The first, called the **direct diagram**, is just a plot of data in  $R^2$  (the observation points with the lines going through pairs of observations). The second diagram, called an **Oja rank plot**, describes data items and regions having particular Oja rank values. From it, quantiles in the Oja sense can be read off. The Oja rank plot is developed in Section 4. In certain senses, the two plots have a duality relationship. Such connections, and other properties, are listed in Section 4. The correspondences between the two plots indicate considerable potential for higher dimensional versions to be useful in informal data analysis.

## 2 Univariate ranks

Given univariate data  $x_1, \dots, x_n$ , the median  $m$  is defined as the choice  $\theta = m$  to minimize the dispersion criterion

$$S(\theta) = \sum_{i=1}^n |x_i - \theta|.$$

Clearly  $S'(\theta) = -\sum \text{sgn}(x_i - \theta)$ , and this gradient function can be used to define univariate rank. In order to facilitate a clear analog with the coming bivariate case it is convenient to define  $\text{sgn}(t) = +1$  if  $t > 0$ ,  $-1$  if  $t < 0$ , but  $\text{sgn}(t)$  can take any value in  $[-1, +1]$  when  $t = 0$ . If  $x_{(j)}$  denotes the  $j$ th order statistic, then  $\frac{1}{2}S'(x_{(j)})$  is any value in  $[j - 1 - \frac{n}{2}, j - \frac{n}{2}]$  while in general, if  $x_{(j-1)} < \theta < x_{(j)}$ , then

$$\frac{1}{2}S'(\theta) = j - 1 - \frac{n}{2}.$$

The rank of the position  $\theta$  among  $\{x_i\}$  can therefore be defined as

$$R(\theta) = \frac{1}{2}S'(\theta),$$

and  $-\frac{1}{2}n \leq R(\theta) \leq \frac{1}{2}n$ , with  $x_{(0)} = -\infty$  and  $x_{(n+1)} = +\infty$ .

Inversion of the rank function leads to the notion of quantile. For  $0 < p < 1$ , the  $p$ th quantile  $\xi_p$  is the solution  $\theta = \xi_p$  of  $R(\theta) = (2p - 1)(\frac{n}{2})$ . It is

easy to verify that if  $p = (j-1)/n$ , then  $\xi_p$  is any value within  $[x_{(j-1)}, x_{(j)}]$ , while if  $(j-1)/n < p < j/n$ , then  $\xi_p = x_{(j)}$ .

The preceding definitions of univariate rank and quantile will now be extended to the bivariate case, using the Oja dispersion function for a bivariate median. Note now that later it is convenient to define bivariate quantiles in a way which is notationally different from the univariate case.

### 3 Oja bivariate ranks

Now suppose that  $x_1, \dots, x_n$  all  $\in R^2$ . The Oja bivariate median  $m$  is the choice  $\theta = m$  to minimize

$$S(\theta) = \sum_{i < j} A(x_i, x_j, \theta)$$

where  $A(a, b, c)$  is the area of the triangle having vertices  $a$ ,  $b$  and  $c$ . The corresponding gradient function is

$$\nabla S(\theta) = \frac{1}{2} \sum_{i < j} u(x_i, x_j; \theta)$$

where  $u$  is a “repulsion vector”, having magnitude  $|x_i - x_j|$  and direction perpendicular to and away from the chord between  $x_i$  and  $x_j$ , towards  $\theta$ . See Brown and Hettmansperger (1987a) for details. Correspondingly, the Oja rank of  $\theta$  is defined as

$$\begin{aligned} R(\theta) &= \nabla S(\theta), \\ &= \frac{1}{2} \sum_{i < j} u(x_i, x_j; \theta) \end{aligned}$$

Note that ranks  $R(\theta)$  are bivariate vectors, with direction as well as magnitude. The orientation of  $R(\theta)$  among other  $\{R(x_i)\}$  will be roughly similar to that of  $\theta$  among  $\{x_i\}$ , but in a general sense the ranks display more regularity than the original data, resembling the situation for univariate ranks.

An important observation is that  $R(\theta)$  remains constant as  $\theta$  changes locally. Furthermore,  $R$  changes value only when  $\theta$  crosses a line connecting some  $x_i, x_j$ . Then the increment to  $R(\theta)$  is  $\pm u(x_i, x_j; \theta)$ . This observation establishes the fundamental basic relationship between the direct diagram, the plot of data  $\{x_i\}$ , and the Oja rank plot, of the rank values. It can be described as follows.

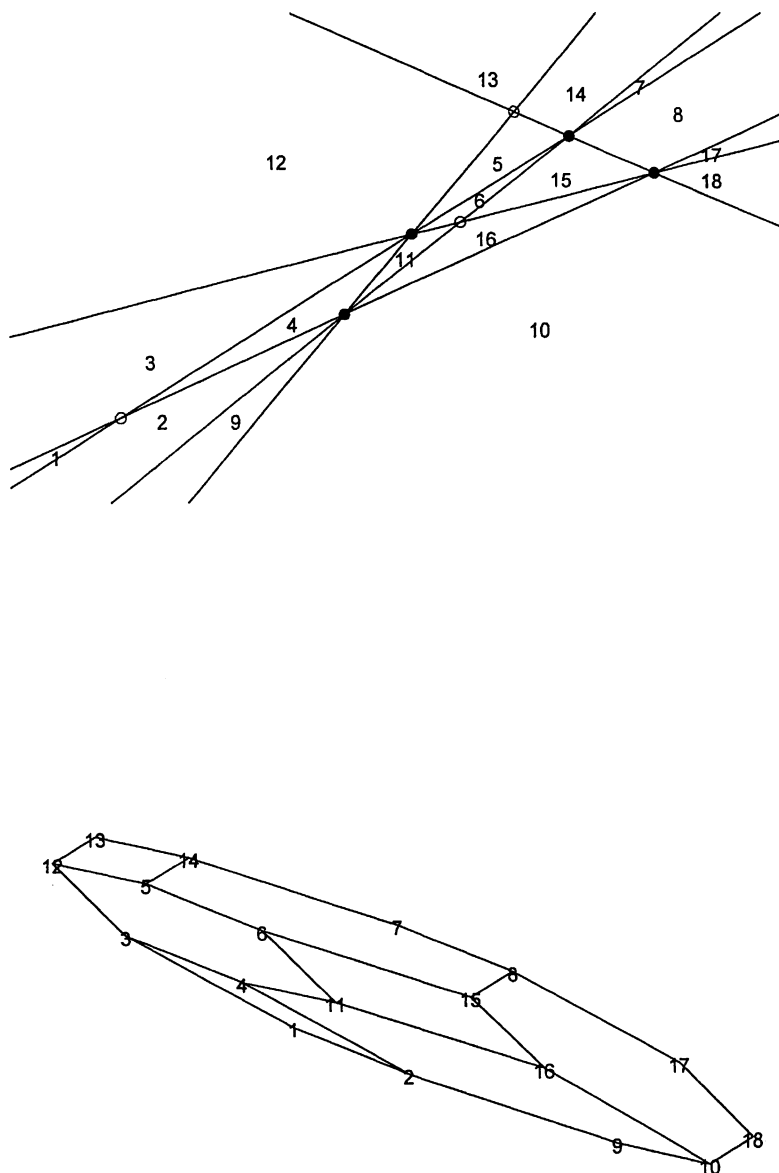


Figure 1: Direct diagram (upper case): Black points are original four data points; white points are secondary points. Tiles are numbered. Rank plot (lower case): Numbered vertices are rank plots of corresponding tiles.

Drawing in extended lines connecting all observation pairs  $x_i, x_j$  defines a **natural tiling** in the direct diagram. The tiles are polygons with vertices at observations, or at intersections of a line connecting  $x_i, x_j$  with a line connecting  $x_k, x_l$ , for some  $i, j, k, l$ . These intersections are called **secondary points**. For  $n$  original data points in the general position (no parallel lines), there are  $3\binom{n}{4} = O(n^4/8)$  secondary points. For the points not in the general position, the number of secondary points (and consequently the number of tiles) is smaller.

The ranks for all  $\theta$  within a tile of the direct diagram are constant vectors. That constant is a point of the Oja rank plot, so rank plot points correspond to tiles in the direct diagram. Neighbouring tiles have rank values differing by a repulsion vector  $u$ . Therefore the rank value of any point of a boundary between tiles can be any point in the rank plot on the chord between the rank value points of the two tiles.

Furthermore, in the direct diagram,  $n$  observation points and  $3\binom{n}{4}$  secondary points lie at the junction of several tiles. Correspondingly, their rank values are not unique, but any value in the polygonal region of the rank plot whose vertices are the ranks of the abutting tiles. For illustration, see Section 4 and in particular Figure 1.

These observations lead to a number of further relationships between direct diagram and rank plot, described in the next section.

## 4 Direct diagram and rank plots

### 4.1 Relations between the tiles

The regions of the rank plot are more regular than the tiles in the direct diagram. Figure 1 show a case of  $n = 4$  data points with just 3 secondary points. A small number of points minimizing the clutter in the figure was used to illustrate the duality relationship in a simple case. See Figure 3 for the rank plot of 10 points.

In general, every data point has  $n - 1$  lines emanating from it, towards other data points, so in the rank plot the rank region for a data point is a  $2(n - 1)$ -sided polygon whose opposite sides are parallel and of equal length, i.e. an order- $(n - 1)$  parallelogram. By contrast, the rank regions of secondary points are conventional 4-sided parallelograms. Together, the two types of parallelogram form an orderly tiling of  $R^2$  in the rank plot. All sides of regions are repulsion vectors as occurring in the definition of Oja rank. Figure 2 provide illustrations.



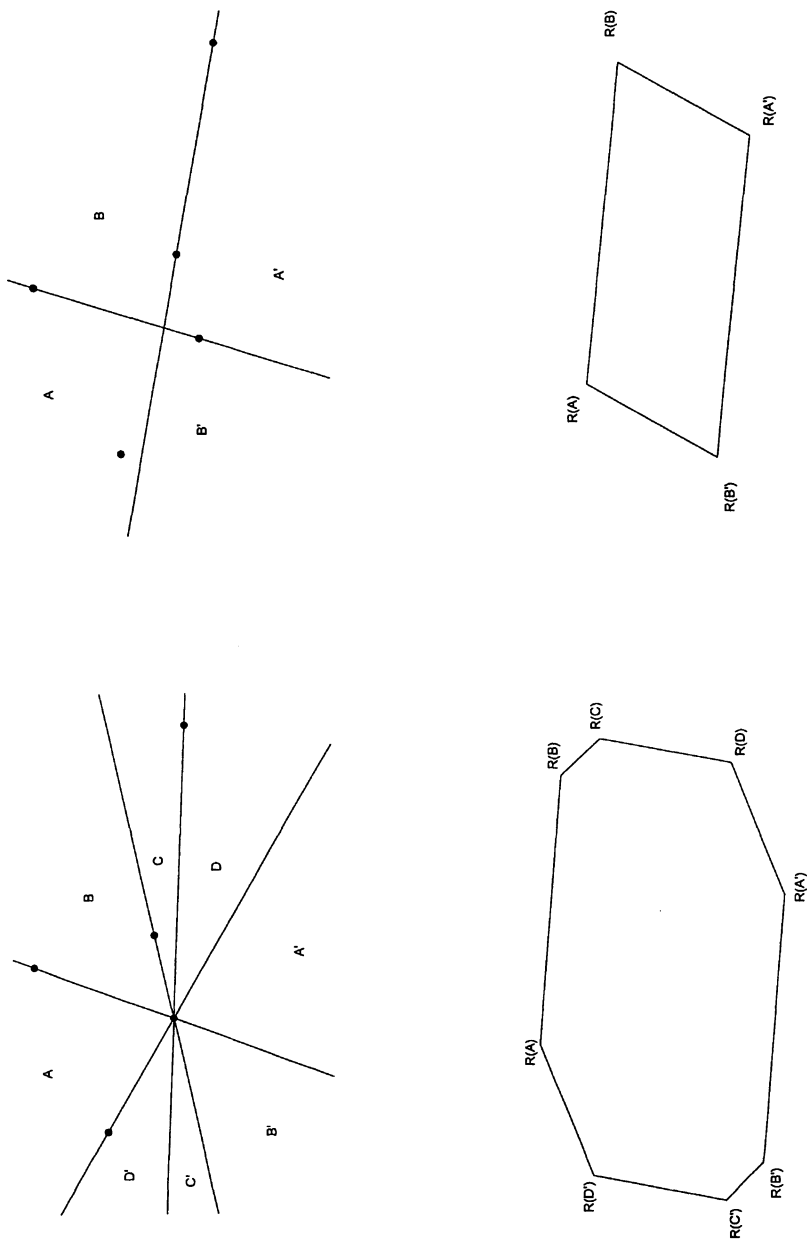


Figure 2: Direct diagrams and rank plots near a data point (lower case) and near a secondary data point (upper case).

There is complete projective duality between the two plots; that is, points in the direct diagram are associated with regions in the rank plot, chords in the direct diagram are associated with chords in the rank plot, and regions in the direct diagram are associated with points in the rank plot:

| Direct plot     | Rank plot                     |
|-----------------|-------------------------------|
| point           | tile: $(n - 1)$ parallelogram |
| secondary point | tile: 2 parallelogram         |
| chord           | chord                         |
| extended line   | set of chords                 |
| tile            | point                         |

Only the location information is lost in the rank plot: The original data can be recaptured from the rank plot and the value of any data point (or Oja median). Let  $B$  be the sample covariance matrix computed on the Oja ranks. The **standardized rank plot** is then obtained if the rank plot items are multiplied (from the left) by  $B^{-1/2}$ . Both location and scale information is lost in the standardized rank plot.

Note that the Oja rank vectors are location invariant and affine equivariant in the sense that if the original observation vectors are multiplied by a full rank matrix  $A$ , the rank plot items will be multiplied by  $A^* = \text{abs}(\det(A))(A^{-1})^T$ . If  $A$  is orthogonal then  $A^* = A$  and if  $A = \text{diag}(a_1, a_2)$  then  $A^* = \text{diag}(a_2, a_1)$ . For elliptical distributions, the eigenvectors from the Oja rank covariance matrix are then the eigenvectors of the conventional covariance matrix but the eigenvalues are reversed. The fact is connected to rank plot scale elongation occurring in orthogonal directions to scale elongations in the direct diagram; see Section 4.3.

## 4.2 The rank plot boundary

In using a rank value to assess the position of a point among points in a data cloud, it is useful to know the extremities of the rank plot. The rank plots are not standardized and the rank plot boundaries are determined by the data. The boundary tiles in the direct diagram each have an open face extending to  $\infty$ , and the rank values of these tiles form the vertices of the convex hull of the rank plot. Plotting these vertices will delineate the rank plot boundary, but there is a quicker more informal method of describing approximately where this boundary is.

This method, yielding an approximate boundary of the rank plot, is as follows.

Consider  $\theta$  far away from the original data cloud, in direction  $\alpha$ , i.e.

approximately

$$\theta = \hat{\theta} + r \begin{pmatrix} \cos \alpha \\ \sin \alpha \end{pmatrix}$$

with  $r \rightarrow \infty$ , where  $\hat{\theta}$  is the Oja median. For  $r$  large,  $R(\theta)$  does not depend on  $r$ . Then the contribution to  $R(\theta)$  in direction  $\alpha$  of the vector  $(1/2)u(x_i, x_j; \theta)$  is approximately of magnitude

$$\frac{1}{2}r_{ij}|\sin(\alpha - \alpha_{ij})|, \quad (1)$$

where  $r_{ij} = |x_i - x_j|$  and the line joining  $x_i, x_j$  has direction  $\alpha_{ij}$  (or  $\alpha_{ij} + \pi$ ). The sum of terms like (1) is awkward because of the absolute value, but a smooth approximation comes from using

$$|\sin x| \simeq a + \frac{1}{2}(1-a)(1 - \cos 2x), \quad -\pi < x < \pi \quad (2)$$

Any value of  $a$  with  $0 < a < 1$  may be used; the error of approximation varies between  $a$  at  $x = 0$  and  $(-1/2)(1 - 2a)^2/(1 - a)$  when  $|\sin x| = (1/2)(1 - a)^{-1}$ . The minimax error is  $(2 - 2^{1/2})/4 = 0.1464$  at  $a = (2 - 2^{1/2})/4 = 0.1464$  and this is a convenient choice for  $a$ . Then summing over terms in (1), using (2), gives

$$\begin{aligned} R(\theta) &= \frac{1}{4}(1+a) \sum_{i < j} r_{ij} - \frac{1}{4}(1-a)\gamma \cos(2\alpha - 2\omega), \\ &= \frac{1}{4}(1+a)R - \frac{1}{4}(1-a)\gamma [\cos(2\alpha) \cos(2\omega) - \sin(2\alpha) \sin(2\omega)], \end{aligned}$$

where

$$\begin{aligned} R &= \sum_{i < j} r_{ij}, \\ \gamma^2 &= \left( \sum_{i < j} r_{ij} \cos(2\alpha_{ij}) \right)^2 + \left( \sum_{i < j} r_{ij} \sin(2\alpha_{ij}) \right)^2, \\ \cos(2\omega) &= \gamma^{-1} \sum r_{ij} \cos(2\alpha_{ij}) \end{aligned}$$

and

$$\sin(2\omega) = \gamma^{-1} \sum r_{ij} \sin(2\alpha_{ij}).$$

The parameters  $\cos(2\omega)$ ,  $\sin(2\omega)$ ,  $\omega$ ,  $\gamma$  and  $R$  are easy to calculate and along with  $a$ , describe  $R(\theta)$  as a simple cosine function of  $\alpha$ , minimum at  $\alpha = \omega \pm \pi$  and maximum at  $\alpha = \omega$ . The corresponding shape of the approximate rank plot boundary is approximately an ellipse whose major and minor axes give a rough indication of the principal components of the

cloud of Oja ranks. See Figure 3 for the rank plot of a data set of 10 observations.

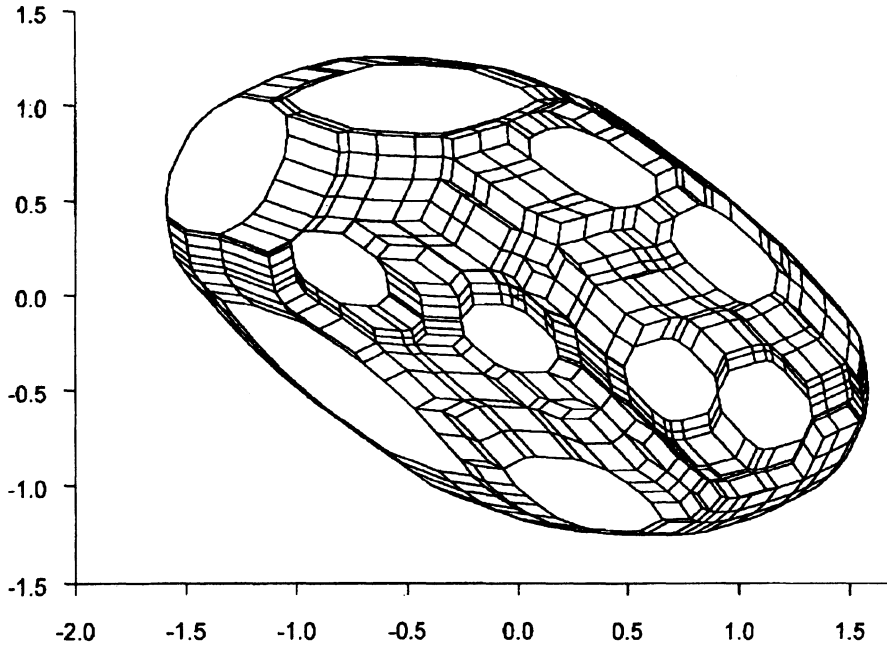


Figure 3: Rank plot for a data set of ten bivariate observations.

Note that

$$\lim_{\theta_1 \rightarrow \infty} R(\theta) = \frac{1}{2} \left( \begin{array}{c} \sum_{i < j} |x_{j2} - x_{i2}| \\ - \sum_{i < j} \text{sign}(x_{j2} - x_{i2})(x_{j1} - x_{i1}) \end{array} \right).$$

If  $x_1, \dots, x_n$  are i.i.d. from a spherical bivariate distribution with marginal Gini mean differences  $\tau = E(|x_{11} - x_{21}|) = E(|x_{12} - x_{22}|)$  then clearly

$$\lim_{\theta_1 \rightarrow \infty} \binom{n}{2}^{-1} R(\theta) \rightarrow_P \frac{1}{2} \begin{pmatrix} \tau \\ 0 \end{pmatrix}$$

and, for

$$\theta = \theta_0 + r \begin{pmatrix} \cos \alpha \\ \sin \alpha \end{pmatrix},$$

$$\lim_{r \rightarrow \infty} \binom{n}{2}^{-1} R(\theta) \rightarrow_P \frac{\tau}{2} \begin{pmatrix} \cos \alpha \\ \sin \alpha \end{pmatrix}.$$

The approximative boundary then is the sphere

$$\left\{ \frac{\tau}{2} u : u^T u = 1 \right\}.$$

For observations from an elliptical distribution

$$PCx_1, \dots, PCx_n,$$

where  $P$  is orthogonal and  $C = \text{diag}(c1, c2)$  diagonal, the asymptotic boundary is then the ellipse

$$\left\{ \frac{\tau}{2} PC^* u : u^T u = 1 \right\},$$

where  $C^* = \text{diag}(c2, c1)$ . Major and minor axes give the principal components for the original bivariate distribution.

### 4.3 Slopes of rank regions

In the rank plot, the rank region of a data point is the order  $(n - 1)$  parallelogram of rank values attributable to that point, whose vertices are the rank values of all the tiles in the direct diagram which abut at the point. There is considerable information in the shape of a rank region as to the position of a data point within a data cloud : The lengths and directions of the  $n - 1$  chords surrounding the rank region give the distances and directions of the other  $n - 1$  points in the direct plot. (The direction is perpendicular  $(\pm \frac{\pi}{2})$  to the direction of the chord.) Figure 4 illustrates how the rank region of an outlier will be elongated in a direction perpendicular to the direction of the rest of the data.

If a rank region has sides predominantly of one direction, the rest of the data is mostly oriented in a perpendicular direction from the data point in question. The lengths of the sides of a rank region are proportional to distances to other data points. Thus an outlier is distinguished by having a rank region with long sides, all with a similar direction.

Other remarks can be made.

(i) If all rank regions tend to have sides of similar direction, then the whole data cloud is elongated in a perpendicular direction.

(ii) Data points towards the center of a data cloud tend to have rank regions whose sides are of mixed lengths. The directions will reflect the general orientation of the data cloud.

(iii) Other data cloud patterns will have corresponding rank plot features. For instance, two separated mini-clouds yield a rank plot whose rank regions have sides tending to be distinctly short, in assorted directions, or distinctly long, in a definite direction perpendicular to the direction between the mini clouds.

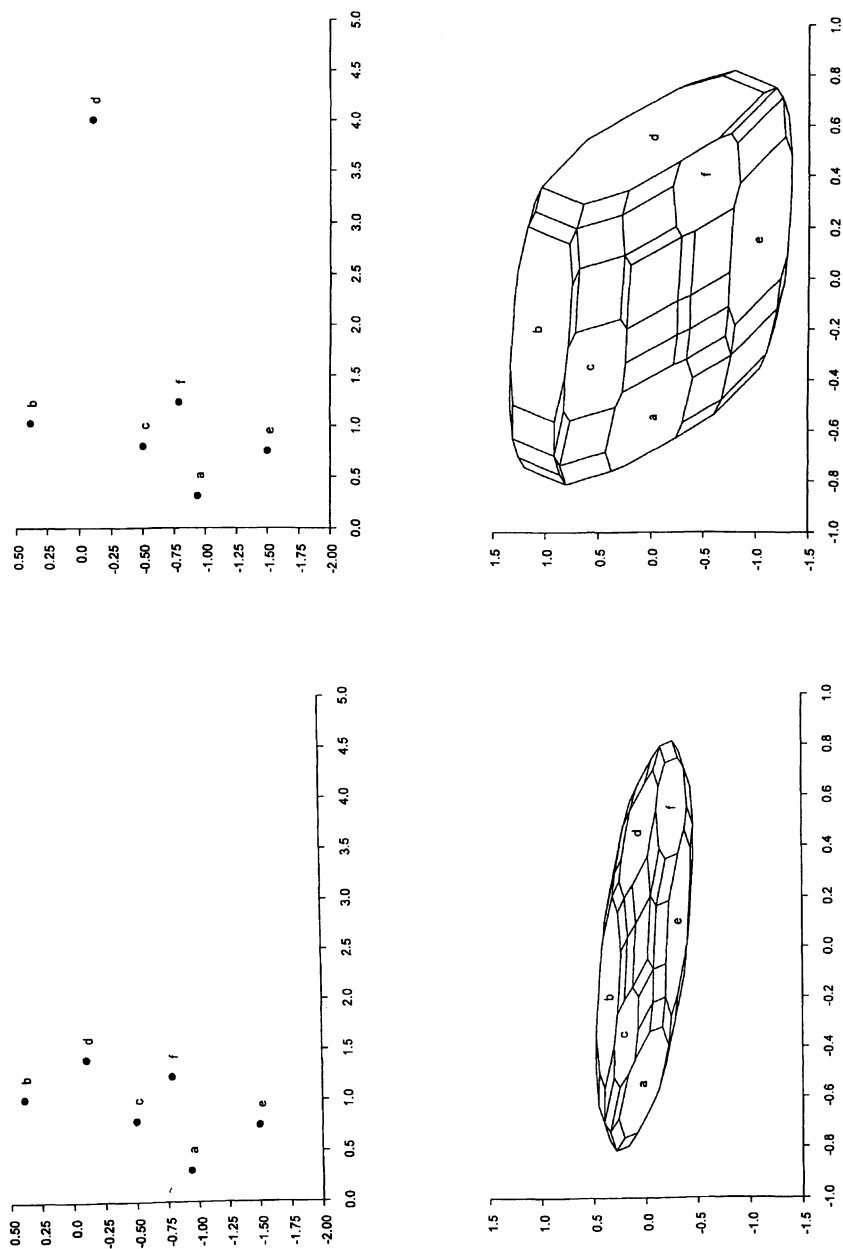


Figure 4: Direct diagrams and data plots for six observations: In the upper case observation  $d$  is moved to be an outlier.

## References

- [1] Brown, B. M. and Hettmansperger, T. P. (1987a). Affine invariant rank methods in the bivariate location model. *J. R. Statist. Soc. B* **49**, 301–310.
- [2] Brown, B. M. and Hettmansperger, T. P. (1987b). Invariant tests in bivariate models and the  $L_1$  criterion. In *Statistical data analysis based on the  $L_1$ -norm and related methods*, Ed. Y. Dodge, pp. 333–344. Amsterdam: North Holland.
- [3] Hettmansperger, T. P., Möttönen, J. and Oja, H. (1997a). Affine invariant multivariate one-sample signed-rank tests. *J. Am. Statist. Assoc.* To appear.
- [4] Hettmansperger, T. P., Möttönen, J. and Oja, H. (1997b). Affine invariant multivariate rank tests for several samples. *Statistica Sinica*. Conditionally accepted.
- [5] Hettmansperger, T. P., Nyblom, J. and Oja, H. (1992). On multivariate notions of sign and rank. In  *$L_1$ -statistical Analysis and Related Methods*, Ed Y. Dodge, pp. 267–278. Amsterdam: North-Holland.
- [6] Liu, R.Y. (1990). On a notion of data depth based upon random simplices. *Ann. Statist.* **18**, 405–414.
- [7] Liu, R.Y. (1992). Data depth and multivariate rank tests. In  *$L_1$ -Statistical Analysis and Related Methods*, Ed. Y. Dodge, pp. 279–302. Amsterdam: North-Holland.
- [8] Liu, R.Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Am. Statist. Ass.* **88**, 252–260.
- [9] Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparam. Statist.* **5**, 201–213.
- [10] Niinimaa, A. and Oja, H. (1997). Multivariate median. In *Encyclopedia of Statistical Sciences*. To appear.
- [11] Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1**, 327–332.
- [12] Small, G. (1990). A survey of multidimensional medians. *Inter. Statist. Rev.* **58**, 263–277.

# Target estimation and implications to robustness

**Luisa Turrin Fernholz**

*Temple University, Princeton, USA*

*Abstract:* This paper considers target functionals  $\tilde{T}$ , which are bias-reduced functionals that can be obtained from a functional  $T$  in a parametric setting. It is shown that the  $L_1$ -error of the corresponding target estimator decreases and the asymptotic normality is obtained using von Mises expansions with the Hadamard derivative. It is also shown that targeting can improve robustness since the gross-error sensitivity decreases under certain conditions. Applications to M-estimates of location, the sample median, and simultaneous M-estimates of location and scale are given.

*Key words:* Bias, influence function,  $L_1$ -error, parametric family, statistical functionals, target estimates, von Mises expansions.

AMS subject classification: Primary 62G99; secondary 62E20.

## 1 Introduction

Target estimation was introduced by Cabrera and Fernholz (1996) as a procedure to reduce the bias and the variance of an estimator. In that paper, the von Mises expansion of a statistical functional was used to obtain conditions for this reduction in bias and variance. Moreover, it was also shown that the bias of the target estimator can be expressed in terms of the influence function of the original statistic. It is natural to ask what the relationship is between the von Mises expansion of the original functional and that of the target functional. This issue is addressed in the present paper where the asymptotic distribution of the target functional will follow from the corresponding von Mises expansion. The influence function of the target functional is also obtained and a condition is given for reducing the



gross-error sensitivity.

The paper is organized as follows: In Section 2 we give a review of the basic ideas behind target functionals and we consider the  $L_1$  error and the mean square error of these functionals. In Section 3 we look at the asymptotics and robustness of target functionals through the von Mises expansion and the influence function. Applications to M-estimates of location, the sample median, and simultaneous M-estimates of location and scale are given in Section 4.

Throughout this paper we shall assume that  $T$  is a statistical functional and the statistic  $T(F_n)$  estimates the parameter  $T(F_\theta)$ , where  $F_n$  is the empirical d.f. corresponding to the sample  $X_1, \dots, X_n$  of i.i.d. random variables with common d.f.  $F_\theta$ , with  $\theta \in \Theta$ , for  $\Theta$  an open subset of the real numbers. When a statistical functional  $T$  satisfies  $T(F_\theta) = \theta$ , the functional is said to be Fisher consistent.

We shall also assume that the expectation of  $T(F_n)$ ,  $g_n(\theta) = E_\theta(T(F_n))$ , exists for all  $\theta \in \Theta$ , where  $E_\theta$  indicates the expectation with respect to  $F_\theta$ . Moreover the function  $g_n$  will be assumed to be one-to-one and differentiable.

## 2 Target functionals

**Definition 1** *Let  $g_n(\theta) = E_\theta(T(F_n))$  be a one-to-one function. The functional  $\tilde{T}_n$  induced by  $T$  from the relation*

$$g_n^{-1}(T) = \tilde{T}_n \quad (1)$$

*will be called the target functional of  $T$ . The statistic  $\tilde{T}_n(F_n)$  will be called the target estimator .*

The above definition was introduced by Cabrera and Fernholz (1996) where the goal was the reduction of the bias and the variance of an estimator. Note that when  $g_n(\theta) = a\theta + b$  for  $a \neq 0$ , then the corresponding target functional is  $\tilde{T}_n = (T - b)/a$ , which will always be unbiased and its variance will be reduced if and only if  $a^2 > 1$ . The variance of  $\tilde{T}_n$  will remain unchanged when  $|a| = 1$ .

It should be noted that Rousseeuw and Ronchetti (1981) used the function  $g_n$  to generate functionals  $g_n^{-1}(T)$  in an entirely different context to study the influence curve of statistics used in testing hypotheses.

The following two theorems refer to the reduction in the bias  $B_{\tilde{T}}(\theta)$  and mean square error of target functionals.

**Theorem 1** *If for a statistical functional  $T$  the function  $g_n(\theta) = E_\theta(T(F_n))$  satisfies*

$$\begin{aligned} (i) \quad & \theta < g_n(\theta) \\ (ii) \quad & 1 < g'_n(\theta) \leq b \\ (iii) \quad & 0 \leq g''_n(\theta) \end{aligned}$$

*then*

$$a) \ E_\theta(\tilde{T}_n) \leq \theta \leq E_\theta(T) \quad \text{and} \quad b) \ |(B_{\tilde{T}_n}(\theta))| < |(B_T(\theta))|.$$

**Theorem 2** *If  $T$  is a statistical functional with variance  $V_T$  and  $g_n(\theta) = E_\theta(T(F_n))$  is differentiable with  $|g'_n(\theta)| > 1$  for all  $\theta \in \Theta$ , then the mean square error of  $\tilde{T}_n$  satisfies*

$$MSE_{\tilde{T}_n} < V_T.$$

The proofs of the above theorems can be found in Cabrera and Fernholz (1996). The other theorems in the current paper are new.

The following result refers to the  $L_1$ -error of the target estimator  $\tilde{T}_n$ .

**Theorem 3** *If  $T$  is a statistical functional and  $g_n(\theta)$  is differentiable with  $|g'_n(\theta)| > 1$  for all  $\theta \in \Theta$ , then*

$$E_\theta|\tilde{T}_n - \theta| < E_\theta|T - E_\theta(T)|.$$

**Proof:** Using the mean value theorem for  $g_n$  we have

$$g_n(\tilde{T}_n) = g_n(\theta) + (\tilde{T}_n - \theta)g'_n(\xi)$$

for some  $\xi$  between  $\theta$  and  $\tilde{T}_n(F_n)$ . Since by definition  $g_n(\tilde{T}_n) = T$ , we have

$$\tilde{T}_n - \theta = \frac{T - g_n(\theta)}{g'_n(\xi)}.$$

By taking absolute values and their expectation of this last equation, we obtain

$$\begin{aligned} E|\tilde{T}_n - \theta| &= E|1/g'_n(\xi)||T - g_n(\theta)| \\ &< E|T - E(T)|, \end{aligned}$$

and the theorem is proved since  $|g'_n(\theta)| > 1$ .  $\square$

An immediate consequence of this theorem is the following corollary that relates the  $L_1$ -errors of  $T$  and  $\tilde{T}$  with the median,  $M_\theta(T)$ , of the distribution of  $T$ .

**Corollary 4** *Under the conditions of Theorem 3 we have*

$$E|\tilde{T}_n - \theta| < E|T - \theta| + |M_\theta(T) - E_\theta(T)|$$

**Proof:** It follows directly from Theorem 3 since

$$\begin{aligned} E|\tilde{T}_n - \theta| &< E|T - M_\theta(T)| + |M_\theta(T) - E_\theta(T)| \\ &\leq E|T - \theta| + |M_\theta(T) - E_\theta(T)| \end{aligned}$$

by the  $L_1$ -minimizing property of the median.  $\square$

This corollary shows that when  $T$  has a symmetric sampling distribution, the target functional  $\tilde{T}$  has smaller  $L_1$ -error.

### 3 Asymptotics and robustness

The results presented in Section 2 refer to the statistic  $\tilde{T}_n(F_n)$ . But for each target functional  $\tilde{T}_n$  we can consider the statistic  $\tilde{T}_n(F_m)$  where  $n$  and  $m$  do not necessarily coincide. For this purpose, let  $F_\theta$  be a parametric family and  $T$  be a Fisher consistent statistical functional. For each  $n$  and  $g_n(\theta)$  as above, we obtain the sequence of target functionals  $\tilde{T}_n = g_n^{-1}(T)$ , each one of them generating statistics  $\tilde{T}_n(F_m)$  for a sample  $X_1, \dots, X_m$ .

Let  $T$  be a Fisher consistent functional with influence function  $\phi_1$  (see Hampel, 1974). For a sample  $X_1, \dots, X_m$  from  $F_\theta$ , the first order von Mises expansion of  $T(F_m)$  is

$$T(F_m) = \theta + \frac{1}{m} \sum_{i=1}^m \phi_1(X_i) + Rem_m. \quad (2)$$

With Hadamard or Frechet differentiability under certain regularity conditions, we have  $Rem_m = o_P(m^{-\frac{1}{2}})$  and  $T$  is asymptotically normal with variance  $\sigma^2 = E(\phi_1(X))^2$  (see Fernholz, 1983).

The following result gives the von Mises expansion of  $\tilde{T}_n$

**Lemma 5** *Let  $T$  be a statistical functional and for a fixed  $n$  let  $\tilde{T}_n$  be the corresponding target functional. If  $T$  is Hadamard differentiable at  $F_\theta$  with von Mises expansion as in (2), then  $\tilde{T}_n$  is also Hadamard differentiable at  $F_\theta$  with von Mises expansion*

$$\tilde{T}_n(F_m) = g_n^{-1}(\theta) + \frac{1}{m} \sum_{i=1}^m (1/g'_n(\theta)) \phi_1(X_i) + \tilde{Rem}_m. \quad (3)$$

**Proof:** First note that  $\tilde{T}_n(F_\theta) = g_n^{-1}(T(F_\theta)) = g_n^{-1}(\theta)$ . Since  $T$  is Hadamard differentiable, the composition  $g_n^{-1}(T) = \tilde{T}_n$  will be Hadamard differentiable for a differentiable real function  $g_n$  by the chain rule (see Fernholz, 1983). Therefore the influence function of  $\tilde{T}_n$  is  $(1/g'_n(\theta))\phi_1(x)$  and the lemma is proved.  $\square$

In general we would be interested in using target estimates when  $T$  is such that  $g_n(\theta) \neq \theta$ . This clearly implies that when  $T$  is Fisher consistent then  $\tilde{T}_n$  will not be. However,  $\tilde{T}_n$  has less bias than  $T$  when  $g_n$  satisfies certain conditions.

When  $n = m$ , the expansion in (3) provides a linear approximation of the statistic  $\tilde{T}_n(F_n)$  with the influence function given by  $(1/g'_n(\theta))\phi_1(x)$ . If we now compare the influence functions of  $\tilde{T}_n$  and  $T$  we can conclude immediately that, when  $|g'_n(\theta)| > 1$ ,  $\tilde{T}_n$  is more robust than  $T$  in terms of gross-error sensitivity (see Hampel et al., 1986) since we have

$$\sup_x |(1/g'_n(\theta))\phi_1(x)| < \sup_x |\phi_1(x)|.$$

When the function  $g(\theta) = a\theta + b$  is linear,  $\tilde{T} = (T - b)/a$  and the corresponding influence function satisfies  $\tilde{\phi}_1(x) = \phi_1(x)/a$ . The gross-error sensitivity of  $\tilde{T}$  will be smaller than that of  $T$  when  $|a| > 1$ .

The expansion in (3) above is useful to derive the asymptotic normality and efficiency of  $\tilde{T}_n$  as we see in

**Theorem 6** *Let  $T$  and  $\tilde{T}_n$  be as in Lemma 4. If  $T$  is Hadamard differentiable at  $F = F_\theta$  with von Mises expansion as in (2) then, for fixed  $n$  and  $m \rightarrow \infty$  we have :*

a)  $\tilde{T}_n$  is asymptotically normal with

$$\sqrt{m}(\tilde{T}_n(F_m) - \tilde{T}_n(F)) \xrightarrow{\mathcal{D}} N(0, \sigma_n^2)$$

where the asymptotic variance of  $\tilde{T}_n$  is  $\sigma_n^2 = (1/g'_n(\theta))^2 \sigma^2$ ;

b) If  $|g'_n(\theta)| > 1$ ,  $\tilde{T}_n$  is asymptotically more efficient than  $T$ .

**Proof:** Part a) is an immediate consequence of Lemma 5 since the Hadamard differentiability of  $\tilde{T}_n$  implies the asymptotic normality of  $\tilde{T}_n$  (see Fernholz, 1983).

For part b), note that when  $|g'_n(\theta)| > 1$ , the asymptotic variance of  $\tilde{T}_n$ , satisfies

$$\sigma_n^2 = (1/g'_n(\theta))^2 \sigma^2 < \sigma^2. \quad \square$$

Theorem 6 refers to the asymptotic normality of  $\tilde{T}_n(F_m)$  for fixed  $n$  when  $m$  tends to infinity. When  $n = m$  is large, we have

**Theorem 7** *Let  $T$  and  $\tilde{T}_n$  be as in Theorem 6. If  $m = n$  tends to infinity and for all  $\theta$   $\sqrt{n}(g_n(\theta) - \theta) \rightarrow 0$ , and  $\sqrt{n}(1/g'_n(\theta) - 1) \rightarrow 0$ , then for all  $\theta$  it holds that*

$$\sqrt{n}(\tilde{T}_n(F_n) - \theta) \xrightarrow{\mathcal{D}} N(0, \sigma^2).$$

**Proof:** From the expansions (2) and (3) above we obtain

$$\begin{aligned} \sqrt{n}(\tilde{T}_n(F_n) - \theta) &= \sqrt{n}(\tilde{T}_n(F_n) - T(F_n) + T(F_n) - \theta) \\ &= \sqrt{n}(\tilde{T}_n(F_n) - T(F_n)) + \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_1(X_i) + \sqrt{n}Rem_n. \end{aligned} \quad (4)$$

Since  $T$  is Hadamard differentiable, we have

$$\sqrt{n}Rem_n \xrightarrow{P} 0$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_1(X_i) \xrightarrow{\mathcal{D}} N(0, \sigma^2).$$

Hence, it suffices to show that the first term in (4) converges to zero in probability. But in

$$\begin{aligned} \sqrt{n}(\tilde{T}_n(F_n) - T(F_n)) &= \sqrt{n}(g_n^{-1}(\theta) - \theta) + \frac{1}{\sqrt{n}} \sum_{i=1}^n (1/g'_n(\theta) - 1)\phi_1(X_i) \\ &\quad + \sqrt{n}(\tilde{Rem}_n - Rem_n) \end{aligned} \quad (5)$$

the first term tends to zero by hypothesis and the third term converges to zero by the Hadamard differentiability of  $T$  and  $\tilde{T}_n$ . For the second term in (5), Markov's inequality implies that for any  $\epsilon > 0$

$$P\left\{\left|\frac{1}{\sqrt{n}} \sum_{i=1}^n (1/g'_n(\theta) - 1)\phi_1(X_i)\right| > \epsilon\right\} \geq \sqrt{n}|1/g'_n(\theta) - 1| E|\phi_1(X)|/\epsilon,$$

which tends to zero when  $n$  tends to infinity since by hypothesis  $\sqrt{n}(1/g'_n(\theta) - 1) \rightarrow 0$  and the theorem is proved.  $\square$

From Theorems 6 and 7 we can see that, when  $n = m$  is large, there is little gain for the target estimators in terms of gross-error sensitivity. However these theorems insure that there will be no loss in gross-error sensitivity when we use targeting to reduce bias.

## 4 Examples

### 4.1 M-estimates of location

Since in this section the sample size  $n$  will always be fixed, we shall simplify the notation by using  $g(\theta)$  instead of  $g_n(\theta)$  and  $\tilde{T}$  for the target functionals. For a parametric family  $F_\theta$ , let the functional  $T(F_\theta) = \hat{\theta}$  be defined implicitly as a solution of

$$\int \psi(x - \hat{\theta}) dF_\theta(x) = 0.$$

The corresponding statistic  $T(F_n)$  is the well known M-estimate of location. When the parametric family satisfies  $F_\theta(x) = F(x - \theta)$  for all  $\theta$  and some d.f.  $F$ , the corresponding M-estimate is location equivariant, that is  $T(F_\theta) = \theta + T(F)$ . It was shown in Cabrera and Fernholz (1996) that the bias of an M-estimate of location is constant. So  $g(\theta) = \theta + B$ , and the influence functions of  $T$  and  $\tilde{T}$  coincide.

### 4.2 The sample median

The functional  $T(F_\theta) = F_\theta^{-1}(1/2) = \hat{\theta}$  corresponds to the sample median  $T(F_n) = F_n^{-1}(1/2)$ . When  $F_\theta$  is not symmetric about  $\theta$  the sample median is biased. The second order von Mises expansion of  $T$  (see Fernholz, 1996; Fankhauser, 1996) gives

$$\begin{aligned} g(\theta) &= E_\theta(F_\theta^{-1}(1/2)) \\ &= \theta - \frac{f'(\theta)}{8n f^3(\theta)} + o(1/n). \end{aligned}$$

Its derivative is

$$g'(\theta) = 1 + \frac{(f'(\theta))^2 - f(\theta)f''(\theta)}{2n f^4(\theta)} + o(1/n),$$

hence for  $\theta$  such that  $f''(\theta) < 0$ , we have  $g'(\theta) > 1$ . In this case  $T$  satisfies the hypotheses of most of the theorems presented above. The target estimator  $g^{-1}(F_n^{-1}(1/2))$  will have smaller bias, MSE, and gross-error sensitivity than  $F_n^{-1}(1/2)$ .

### 4.3 Simultaneous M-estimates of location and scale.

Consider a family of d.f.'s  $F_\theta$  with  $\theta = (\mu, \sigma)$  and the two-dimensional functional  $T(F_\theta) = (T_1(F_\theta), T_2(F_\theta))$  defined implicitly by

$$\int \psi\left(\frac{x - T_1(F_\theta)}{T_2(F_\theta)}\right) dF_\theta(x) = 0,$$

where  $\psi = (\psi_1, \psi_2)$ . The corresponding statistic  $T(F_n) = (T_1(F_n), T_2(F_n))$  satisfies a system of equations and is called an M-estimate of location and scale. See Huber (1981) and Hampel et al (1986).

When the family of distributions is such that  $F_\theta(x) = F(\frac{x-\mu}{\sigma})$  for some fixed d.f.  $F$ , the functional  $T$  satisfies  $T(F_{\mu,\sigma}) = (\mu + \sigma T_1(F), \sigma T_2(F))$  and is said to be location-scale equivariant (see Hampel et al, 1986). In this case, it can be shown that

$$g(\theta) = g(\mu, \sigma) = (\mu + \sigma C_1, \sigma C_2)$$

where  $C_1$  and  $C_2$  are constants independent of  $\mu$  and  $\sigma$ . Now, the target functional is  $\tilde{T} = (\tilde{T}_1, \tilde{T}_2)$  with

$$\begin{aligned}\tilde{T}_1 &= T_1 - (C_1/C_2)T_2 \\ \tilde{T}_2 &= (1/C_2)T_2,\end{aligned}$$

and so the corresponding two-dimensional influence function is  $\tilde{\phi} = (\tilde{\phi}_1, \tilde{\phi}_2)$  where

$$\begin{aligned}\tilde{\phi}_1(x) &= \phi_1(x) - (C_1/C_2)\phi_2(x) \\ \tilde{\phi}_2(x) &= (1/C_2)\phi_2(x).\end{aligned}$$

If we let  $\| \cdot \|$  denote the Euclidean norm, then for each  $x$

$$\begin{aligned}\|\tilde{\phi}(x)\|^2 &= |\tilde{\phi}_1(x)|^2 + |\tilde{\phi}_2(x)|^2 \\ &= |\phi_1(x) - (C_1/C_2)\phi_2(x)|^2 + |(1/C_2)\phi_2(x)|^2 \\ &\leq |\phi_1(x)|^2 + |(C_1/C_2)\phi_2(x)|^2 + |(1/C_2)\phi_2(x)|^2 \\ &= |\phi_1(x)|^2 + ((C_1/C_2)^2 + (1/C_2)^2)|\phi_2(x)|^2 \\ &< |\phi_1(x)|^2 + |\phi_2(x)|^2 \\ &= \|\phi(x)\|^2\end{aligned}$$

when  $C_1^2 + 1 < C_2^2$ . Therefore the gross-error sensitivity of  $\tilde{T}$  will be smaller than that of  $T$  when  $C_1^2 + 1 < C_2^2$ .

## 5 Closing remarks.

We should note that the regularity conditions of the theorems in Sections 2 and 3 are not too stringent as shown by the examples and applications of target estimation presented in Section 4 as well as in Cabrera and Fernholz (1996). Moreover, the hypotheses of Theorem 7 are not as restrictive as they may seem. The condition  $(g_n(\theta) - \theta) = o(1/\sqrt{n})$  is equivalent to

$E_\theta(Rem_n) = o(1/\sqrt{n})$  which is quite plausible since  $Rem_n = o_P(1/\sqrt{n})$  for Hadamard differentiable functionals. The condition referring to the derivative of  $g_n$  is also satisfied by the most reasonable statistics. The examples presented in Section 4 all satisfy these conditions.

Target estimation is a computer intensive procedure that has proved to be very effective in reducing the bias and the variance of estimators. Target estimation can also be performed when the function  $g_n(\theta)$  is the median of the statistic, as was first presented in Cabrera and Watson (1997). The examples given in Cabrera and Fernholz (1996) and Cabrera and Watson (1997), as well as the applications to practical problems in computer vision of Cabrera and Meer (1996), reveal that target estimation is an effective method of bias and variance reduction in many situations. This paper shows the additional gain in robustness due to targeting.

Although the problems of bias reduction and robustness of an estimator seem to be two independent issues, we showed in this paper that the conditions for bias and variance reduction will assure a smaller bound for the influence function of the bias-reduced functional, which means smaller gross-error sensitivity for the target estimator. The von Mises expansions proved to be a powerful tool to approach the theoretical issues of target estimation. Simulations and practical applications need to be performed now to evaluate more precisely the gain in robustness when estimators are targeted. This is a topic of ongoing research.

## References

- [1] Cabrera, J. and Fernholz, L.T. (1996). Target estimation for bias and variance reduction. Submitted for publication.
- [2] Cabrera, J. and Meer, P. (1996). *Unbiased Estimation of Ellipses by Bootstrapping*. In Press.
- [3] Cabrera, J. and Watson, G.S. (1997). Simulation methods for mean and median bias reduction in parametric estimation. *J. Statist. Plann. Infer.* **57** 143-152
- [4] Fankhauser, E. (1996). Etude du développement de von Mises au deuxième ordre. Ph.D. dissertation. Ecole Polytechnique Federale, Lausanne, Switzerland.
- [5] Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Lecture Notes in Statistics Vol. 19. New York: Springer Verlag.
- [6] Fernholz, L.T. (1996). On higher order von Mises expansions. Technical Report 96-04, Temple University, Dept. Statistics.
- [7] Hampel, F. (1974). The influence curve and its role in robust estima-



- tion. *J. Am. Statis. Assoc.* **69** 383-393.
- [8] Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W. (1986). *Robust Statistics. The Approach Based on the Influence Function*. New York: Wiley.
- [9] Huber, P. (1981). *Robust Statistics*. New York: Wiley.
- [10] Rousseeuw, P. and Ronchetti, E. (1981). Influence curves of general statistics. *Journal of Computational and Applied Mathematics* **7** 161-166.

# Model checks in statistics: An innovation process approach

Winfried Stute

*University of Giessen, Germany*

*Abstract:* In this paper we study a class of Gaussian processes which typically appear as limits of marked empirical processes when composite models need to be checked. A transformation to their martingale part is derived which when applied to the empirical process gives rise to asymptotically distribution-free tests for composite models.

*Key words:* Model checks, innovation martingale, gaussian process.

AMS subject classification: 62G10, 62G30, 60G15.

## 1 Introduction

In this paper we will develop a general methodology for nonparametrically testing the goodness-of-fit of a parametric or a semiparametric model. To begin with the simplest example, assume one observes independent identically distributed (i.i.d.) random variables  $X_1, \dots, X_n$  on the real line, from some unknown distribution function (d.f.)  $F$ . Furthermore, let  $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$  be a given family of distribution functions parametrized by some vector  $\theta \in \Theta \subset \mathbb{R}^k$ . To keep the discussion as simple as possible, we will assume that no nuisance parameters are present so that  $F_\theta$  is uniquely determined by  $\theta$ . The problem of how to test for the hypothesis

$$H_0 : F \in \mathcal{F}$$

has attracted many researchers over the past decades. Most of the test statistics are certain functionals of the underlying empirical process. More

precisely, denote with

$$F_n(x) = n^{-1} \sum_{i=1}^n 1_{\{X_i \leq x\}}, \quad x \in \mathbb{R},$$

the empirical distribution function of the data. The by now classical invariance principle of Donsker (1952) then asserts that the empirical process

$$\alpha_n(x) = n^{1/2}[F_n(x) - F(x)], \quad (1)$$

in the Skorokhod space  $D[-\infty, \infty]$ , converges in distribution to

$$\alpha_\infty := B^0 \circ F.$$

Here,  $B^0$  is a Brownian Bridge on the unit interval, i.e., a centered Gaussian process with covariance function

$$\text{Cov}[B^0(s), B^0(t)] = \min(s, t) - st.$$

For details and extensions, see Gaenssler and Stute (1979) and Shorack and Wellner (1986). To test for a simple hypothesis,  $F = F_{\theta_0}$ , one needs to replace  $F$  in (1) by  $F_{\theta_0}$  so that under  $H_0$

$$\alpha_n^0 \equiv n^{1/2}[F_n - F_{\theta_0}]$$

equals  $\alpha_n$ . In particular, critical values if not available for finite sample size may be obtained from the distribution of the limit  $\alpha_\infty$ . For composite hypotheses, things unfortunately become more complicated. Under  $H_0, F = F_{\theta_0}$  for some unknown  $\theta_0 \in \Theta$ , the true parameter. Since now  $\theta_0$  remains unspecified, it needs to be estimated from the data by some  $\theta_n$ , say. We thus come up with the so-called empirical process with estimated parameters

$$\hat{\alpha}_n \equiv n^{1/2}[F_n - F_{\theta_n}].$$

This process may be viewed as a basic device to measure the deviance between a completely nonparametric and a parametric fit. It has been extensively studied by Durbin (1973). To briefly recall its ingredients, assume that  $\theta_n$  has, under  $H_0$ , a linear expansion

$$n^{1/2}(\theta_n - \theta_0) = n^{-1/2} \sum_{i=1}^n l(X_i, \theta_0) + o_{\mathbb{P}}(1),$$

where  $l$  is a proper vector-valued function with expectation zero and finite covariance matrix. Then, under appropriate smoothness assumptions,

$$\hat{\alpha}_n(x) = \alpha_n(x) - G^t(x, \theta_0) \int l(y, \theta_0) \alpha_n(dy) + o_{\mathbb{P}}(1)$$

uniformly in  $x$ , where

$$G(x) \equiv G(x, \theta_0) = \left. \frac{\partial F_\theta(x)}{\partial \theta} \right|_{\theta=\theta_0}.$$

From this we readily get

$$\hat{\alpha}_n \rightarrow B^0 \circ F - G^t V \equiv \hat{\alpha}_\infty$$

with

$$V = \int l(y, \theta_0) B^0 \circ F(dy).$$

The limit  $\hat{\alpha}_\infty$  is again a centered Gaussian process, but its covariance function is more complicated, and tables for critical values may and will depend on  $\theta_0$  and are not readily available. In such a situation a parametric bootstrap may offer a useful possibility to approximate the distribution of  $\hat{\alpha}_n$  under  $H_0$ ; see Stute et al. (1993).

Though from a computational point of view, this seems to be quite satisfactory, it is worthwhile considering also another approach which not only provides an approximation in distribution, but also leads to a deeper understanding of the involved processes. For  $\hat{\alpha}_n$ , this approach has been initiated, in a landmark paper, by Khmaladze (1981). As to this, recall that  $B^0$  has the representation

$$B^0(t) = B(t) - tB(1), \quad 0 \leq t \leq 1,$$

in terms of a Brownian Motion  $B$  and, vice versa,

$$B(t) = B^0(t) + \int_0^t \frac{B^0(x)}{1-x} dx. \quad (2)$$

In the latter equation  $B$  may be viewed as the innovation martingale and the integral as the compensator in the Doob-Meyer decomposition of  $B^0$ . Now, Khmaladze (1981) was able to also find the corresponding decomposition for  $\hat{\alpha}_\infty$ . Replacing  $\hat{\alpha}_\infty$  by its innovation martingale then leads to a new process, say  $T\hat{\alpha}_\infty$ , which is a Gaussian martingale and hence a Brownian Motion w.r.t. proper time. In particular, this process is distribution-free modulo a transformation in time and therefore is a good candidate for giving rise to distribution-free test statistics.

It is the purpose of the present paper to extend Khmaladze's (1981) approach to a much more general setting. This will enable us to design model checks in the context of regression, times series, multivariate analysis and survival analysis, among others. Now, rather than (2), our starting point

will be the following representation of  $B^0$  in terms of  $B$ , which incorporates a transformation in time and a scale factor:

$$B^0(t) = (1-t)B\left(\frac{t}{1-t}\right). \quad (3)$$

To show that the right hand side has the same covariance structure as  $B^0$ , just use the monotonicity of the time transformation and apply

$$\text{Cov}[B(s), B(t)] = \min(s, t).$$

Monotonicity will also be a crucial issue in the examples which will be shortly discussed. In each case the limit process will be of the following type:

$$R_\infty = G_1 B \circ \psi - G_2^t V. \quad (4)$$

Here,  $G_1$  and  $G_2$  are two deterministic functions,  $\psi$  denotes the aforementioned nondecreasing nonnegative time transformation and  $V$  is a normal vector, which may and will depend on  $B$ . Conclude from the introductory remarks that for  $R_\infty = \hat{\alpha}_\infty$ , i.e., for the empirical process with estimated parameters,

$$G_1 = 1 - F \quad \psi = F/(1 - F)$$

and

$$G_2 = \frac{\partial F_\theta}{\partial \theta} \quad \text{at } \theta = \theta_0.$$

In our second example we discuss a situation which typically comes up when the  $X$ -data represent lifetimes. Under random right censorship one observes, due to other causes of failure, variables  $Z_i = \min(X_i, Y_i)$ ,  $1 \leq i \leq n$ , where the censoring variables are independent and also independent of the  $X$ 's, with the common d.f.  $G$ . Also available are 0-1 variables  $\delta_i = 1_{\{X_i \leq Y_i\}}$  indicating whether  $X_i$  has been observed or not. Since under censorship  $F_n$  may not be available, it needs to be replaced by the nonparametric MLE adapted to the new framework:

$$1 - \hat{F}_n(x) = \prod_{i=1}^n \left[ 1 - \frac{\delta_{[i:n]}}{n - i + 1} \right]^{1_{\{Z_{i:n} \leq x\}}}, \quad x \geq 0. \quad (5)$$

This is the famous product-limit estimator due to Kaplan-Meier (1958). In (5),  $Z_{1:n} \leq \dots \leq Z_{n:n}$  are the order statistics of the observed  $Z$ 's. Finally  $\delta_{[i:n]}$  denotes the  $\delta$ -variable associated with  $Z_{i:n}$ . Note that  $\hat{F}_n$  boils down to  $F_n$  if all  $\delta$ 's equal one. Breslow and Crowley (1974) extended Donsker's invariance principle to the present setup. They showed that the so-called Kaplan-Meier process

$$\beta_n(x) = n^{1/2}[\hat{F}_n(x) - F(x)]$$

converges in distribution to a centered Gaussian process  $\beta_\infty$ . In our notation it admits the representation

$$\beta_\infty = (1 - F)B \circ C,$$

where, under a continuity assumption,

$$C(x) = \int_0^x \frac{F(dy)}{(1 - F(y))^2(1 - G(y))}.$$

Hence, in terms of (4), the Kaplan-Meier process with estimated parameters converges to  $R_\infty$  with

$$G_1 = 1 - F \quad \text{and} \quad \psi = C.$$

The function  $G_2$  is the same as before. A detailed analysis of this example may be found in Nikabadze and Stute (1997).

In our next example, we will discuss the important problem of model checks in regression. For this, let  $(X, Y)$  be a bivariate random vector such that  $\mathbb{E}|Y| < \infty$ . Denote with

$$m(x) = \mathbb{E}\{Y|X = x\}$$

the regression function of  $Y$  w.r.t.  $X = x$ . Also, let  $\mathcal{M} = \{m_\theta : \theta \in \Theta\}$  be a given parametric family of candidates for  $m$ . For example, the  $m_\theta$ 's may consist of all functions spanned by a given basis  $g_1, \dots, g_k$ :

$$m_\theta(x) = \theta_1 g_1(x) + \dots + \theta_k g_k(x).$$

This includes, e.g., all polynomials or trigonometric polynomials with a given bound on the degree. To test for the hypothesis

$$H_0 : m \in \mathcal{M},$$

let  $\theta_n$  be, under  $H_0$ , any estimator of  $\theta_0$ , computed from a sample of independent replicates of  $(X, Y)$ , admitting a representation

$$n^{1/2}(\theta_n - \theta_0) = n^{-1/2} \sum_{i=1}^n l(X_i, Y_i, \theta_0) + o_{\mathbb{P}}(1).$$

The residuals

$$\hat{\varepsilon}_{in} = Y_i - m_{\theta_n}(X_i), \quad 1 \leq i \leq n,$$

traditionally play an important role in model diagnostics for regression. In our approach they will be embedded into a marked point process

$$\hat{\gamma}_n(x) = n^{-1/2} \sum_{i=1}^n \hat{\varepsilon}_{in} 1_{\{X_i \leq x\}}, \quad x \in \mathbb{R}.$$

Under  $H_0$ , one can show that

$$\hat{\gamma}_n \rightarrow \hat{\gamma}_\infty = B \circ \psi - G_2^t V,$$

where

$$\psi(x) = \int_{-\infty}^x \sigma^2(u) F(du)$$

and

$$G_2(x) = \int_{-\infty}^x \frac{\partial m_\theta(u)}{\partial \theta} F(du) \quad \text{at } \theta = \theta_0,$$

with  $\sigma^2(u) = \text{Var}\{Y|X = u\}$  denoting the conditional variance and  $F$  being the marginal distribution of  $X$ . See Stute (1997) and Stute, Thies and Zhu (1996) for details. We thus see that (4) applies again with  $G_1 \equiv 1$  and  $G_2, \psi$  from above.

Another example to which our methodology will apply is in a time series context. For this, let  $X_1, X_2, \dots$  be a stationary sequence of observations. We are interested in the dynamics of the process. One possibility would be to decompose a future observation  $X_i$  into the part explained by the past observations and the  $i$ -th innovation:

$$X_i = m(X_{i-1}, X_{i-2}, \dots) + \varepsilon_i.$$

Thus  $m$  is the regression function of  $X_i$  given  $\mathcal{F}_{i-1} = \sigma(X_{i-1}, X_{i-2}, \dots)$ . If we are, e.g., interested in testing whether the  $X$ -sequence is first order autoregressive with  $m \in \mathcal{M}$ , a pre-specified parametric model, we could form, similar to the regression case, the process

$$\hat{\delta}_n(x) = n^{-1/2} \sum_{i=1}^n [X_i - m_{\theta_n}(X_{i-1})] 1_{\{X_{i-1} \leq x\}}.$$

Due to dependencies some little extra work is needed to show that also in this case  $\hat{\delta}_n \rightarrow \hat{\delta}_\infty$ , where  $\hat{\delta}_\infty$  is of type (4) with  $G_1 = 1$  and some appropriate  $\psi$  and  $G_2$ . Note that the stationary distribution now also depends on  $\theta_0$ . See Koul and Stute (1997) for details.

Our final example concerns a generalized linear model. Here one observes a sequence of multivariate data  $(X_i, Y_i), 1 \leq i \leq n$ , from  $\mathbb{R}^{k+1}$ ,

for which it is assumed that the regression function of  $Y_1$  given  $X_1$  has a decomposition into a linear form of  $X_1$  and a specified link function  $h$ :

$$m(\underline{x}) = \mathbb{E}[Y_1 | X_1 = \underline{x}] = h(\theta_{10}x_1 + \dots + \theta_{k0}x_k).$$

The associated process for testing that  $m$  is of this form becomes

$$\hat{\varepsilon}_n(x) = n^{-1/2} \sum_{i=1}^n [Y_i - h(\langle \theta_n, X_i \rangle)] 1_{\{\langle \theta_n, X_i \rangle \leq x\}},$$

where  $\theta_n$  is an estimator of  $\theta_0 = (\theta_{10}, \dots, \theta_{k0})$  and  $\langle \cdot, \cdot \rangle$  is the scalar product in  $\mathbb{R}^k$ .

Again it can be shown that under standard regularity assumptions  $\hat{\varepsilon}_n$  in the limit is of the form (4). The case when  $h$  is unspecified requires nonparametric estimation of the (univariate) link function.

This list of examples indicates that the class of Gaussian processes considered in (4) is rich enough to cover many interesting cases which typically appear as limit processes when parameters need to be estimated. Since their distributional character is not readily understood, we propose to transform  $R_\infty$  from (4) to another process, which has a much nicer structure, namely a Brownian Motion in proper time. This will be the content of the following section.

## 2 Transformation of Gaussian processes

As we have seen in the first section Gaussian processes of type (4)

$$R_\infty = G_1 B \circ \psi - G_2^t V$$

frequently appear as limits of certain marked empirical processes when parameters need to be estimated. In this section we introduce a transformation  $T$  which maps  $R_\infty$  into a Brownian Motion in proper time. This transformation will be a composition of two linear operators  $T_1$  and  $T_2$  which will be defined now.

Assume that  $G_1$  is a function of bounded variation which is positive on its support. For the sake of simplicity only a continuous  $G_1$  will be considered. Put

$$(T_1 f)(x) = f(x) - \int_{-\infty}^x \frac{f(y)}{G_1(y)} G_1(dy). \quad (6)$$

Here  $f$  varies in the class of functions for which the integral is defined.



**Lemma 1** *The stochastic process  $T_1 G_1 B \circ \psi$  is a Brownian Motion w.r.t. time*

$$\varphi(x) = \int_{-\infty}^x G_1^2(y) \psi(dy).$$

**Proof:** We have

$$T_1 G_1 B \circ \psi(x) = G_1 B \circ \psi(x) - \int_{-\infty}^x B \circ \psi dG_1 = \int_{-\infty}^x G_1 dB \circ \psi.$$

It follows that  $T_1 G_1 B \circ \psi$  is a centered Gaussian process with covariance function  $\min\{\varphi(x_1), \varphi(x_2)\}$  at  $x_1, x_2$ .  $\square$

For the empirical process and the Kaplan-Meier process the function  $G_1$  equals  $1 - F$  so that

$$T_1 f = f + \int \frac{f}{1 - F} dF,$$

which corresponds to (2). For the other examples,  $G_1 \equiv 1$  in which case the integral in (6) vanishes and  $T_1$  reduces to the identity operator.

Since  $T_1$  is a linear operator and since  $V$  does not depend on  $x$ , we obtain

$$\begin{aligned} T_1 R_\infty &= T_1 G_1 B \circ \psi - T_1 G_2^t V \\ &= B \circ \varphi - (T_1 G_2)^t V \equiv B \circ \varphi - G_3^t V, \end{aligned}$$

say, where

$$\begin{aligned} G_3 &= T_1 G_2 = G_2 - \int \frac{G_2}{G_1} dG_1 \\ &= \int \left[ \frac{dG_2}{dG_1} - \frac{G_2}{G_1} \right] dG_1, \end{aligned}$$

provided the Radon-Nikodym derivative of  $G_2$  w.r.t.  $G_1$  exists. In the next step we construct a linear operator  $T_2$  with the following two properties:

$$T_2 G_3 \equiv 0 \tag{7}$$

$$T_2 B \circ \varphi = B \circ \varphi \quad \text{in distribution.} \tag{8}$$

Putting  $T = T_2 \circ T_1$  we therefore get in distribution

$$T R_\infty = T_2(B \circ \varphi - G_3^t V) = B \circ \varphi,$$

i.e.,  $T R_\infty$  is a Brownian Motion w.r.t. time  $\varphi$ .

To define  $T_2$ , let  $R_\infty^0 = B \circ \varphi$  be a Brownian Motion w.r.t. time  $\varphi$ . Also, let  $G$  be a given vector-valued function. Define the matrix

$$A(x) = \int_x^\infty \left( \frac{dG}{d\varphi} \right) \left( \frac{dG}{d\varphi} \right)^t d\varphi$$

and

$$T_2 f(x) = f(x) - \int_{-\infty}^x \left( \frac{dG}{d\varphi} \right)^t (y) A^{-1}(y) \left[ \int_y^\infty \frac{dG}{d\varphi}(z) f(dz) \right] \varphi(dy) \quad (9)$$

assuming that  $A$  is nonsingular.

**Lemma 2** *We have*

$$(i) \quad T_2 G^t \equiv 0$$

$$(ii) \quad T_2 R_\infty^0 = R_\infty^0 \text{ in distribution}$$

**Proof:** (i) is trivial; as to (ii), we have for  $s \leq t$ ,

$$\begin{aligned} \text{Cov}[T_2 R_\infty^0(s), T_2 R_\infty^0(t)] &= E[R_\infty^0(s) R_\infty^0(t)] \\ &- E \left\{ R_\infty^0(t) \int_{-\infty}^s \left( \frac{dG}{d\varphi} \right)^t (y) A^{-1}(y) \left[ \int_y^\infty \frac{dG}{d\varphi}(z) R_\infty^0(dz) \right] \varphi(dy) \right\} \\ &- E \left\{ R_\infty^0(s) \int_{-\infty}^t \left( \frac{dG}{d\varphi} \right)^t (y) A^{-1}(y) \left[ \int_y^\infty \frac{dG}{d\varphi}(z) R_\infty^0(dz) \right] \varphi(dy) \right\} \\ &+ E \left\{ \int_{-\infty}^s \int_{-\infty}^t \left( \frac{dG}{d\varphi} \right)^t (y_1) A^{-1}(y_1) \left[ \int_{y_1}^\infty \frac{dG}{d\varphi}(z) R_\infty^0(dz) \right] \varphi(dy_1) \right. \\ &\quad \left. \left[ \int_{y_2}^\infty \left( \frac{dG}{d\varphi} \right)^t (z) R_\infty^0(dz) \right] A^{-1}(y_2) \left( \frac{dG}{d\varphi} \right) (y_2) \varphi(dy_2) \right\}. \end{aligned}$$

The first expectation equals  $\varphi(s)$ , while the second is easily seen to be

$$\int_{-\infty}^s \left( \frac{dG}{d\varphi} \right)^t (y) A^{-1}(y) \int_y^t \frac{dG}{d\varphi}(z) \varphi(dz) \varphi(dy).$$

Finally, the third and fourth expectations equal

$$\int_{-\infty}^s \left( \frac{dG}{d\varphi} \right)^t (y) A^{-1}(y) \int_y^s \frac{dG}{d\varphi}(z) \varphi(dz) \varphi(dy)$$

and

$$\int_{-\infty}^s \int_{-\infty}^t \left( \frac{dG}{d\varphi} \right)^t (y_1) A^{-1}(y_1) A(y_1 \vee y_2) A^{-1}(y_2) \frac{dG}{d\varphi}(y_2) \varphi(dy_2) \varphi(dy_1),$$

respectively. Summation and an application of Fubini complete the proof.  $\square$

**Theorem 1** *Assume that*

$$R_\infty = G_1 B \circ \psi - G_2^t V.$$

*Define  $T_1$  through (6) and  $T_2$  through (9), with  $G = G_3$ . Then  $T \equiv T_2 \circ T_1$  satisfies*

$$TR_\infty = B \circ \varphi \quad \text{in distribution.}$$

**Proof:** Apply Lemma 1, (7) and (8).  $\square$

We now briefly discuss further issues needed before Theorem 1 can be applied to a real data situation. Let  $R_n$  be one of the processes  $\hat{\alpha}_n - \hat{\varepsilon}_n$  considered in the previous section, or any other marked empirical process admitting a limit  $R_\infty$  as given in (4). The next step to verify is that along with

$$R_n \rightarrow R_\infty$$

one has

$$TR_n \rightarrow TR_\infty = B \circ \varphi. \quad (10)$$

Finally, observe that typically  $T$  incorporates quantities which are unknown in practice and need to be estimated from the data. Hence we come up with a random operator  $T_n$ , for which it remains to show that

$$T_n R_n \rightarrow B \circ \varphi. \quad (11)$$

The proof of (10) and (11) requires some extra work and uses special properties of the underlying processes. It is therefore beyond the scope of the present paper. For the aforementioned examples technical details as well as simulation results may be found in the cited papers.

We finally discuss an application of (11) which is designed to derive tests for  $H_0$  when the alternative is specified. As has been noted by Stute (1997) in the regression case, the Radon-Nikodym derivative of the underlying test process  $R_n$  w.r.t. the hypothesis and local alternatives may often be expressed, in the limit, in terms of the principal components of  $R_\infty$ . While these are not readily available and some numerical work is required

to approximate them, the transformed processes converge to a Brownian Motion, for which the principal components are readily available. In other words, Theorem 1 together with (11) may be used to yield optimal Neyman-Pearson tests for composite models when local alternatives are specified.

## References

- [1] Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimators under random censorship. *Ann. Statist.* **2** 437-453.
- [2] Donsker, M.D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **23** 277-281.
- [3] Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Ann. Statist.* **1** 279-290.
- [4] Gaenssler, P. and Stute, W. (1979). Empirical processes: a survey of results for independent and identically distributed random variables. *Ann. Prob.* **7** 193-243.
- [5] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53** 457-481.
- [6] Khmaladze, E.V. (1981). Martingale approach in the theory of goodness-of-fit tests. *Theor. Prob. Appl.* **26** 240-257.
- [7] Koul, H.L. and Stute, W. (1997). Nonparametric model checks for time series. Preprint. Univ. of Giessen.
- [8] Nikabadze, A. and Stute, W. (1997). Model checks under random censorship. *Statist. and Prob Letters* **32** 249-259.
- [9] Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*. New York: Wiley
- [10] Stute, W., González Manteiga, W. and Presedo Quindimil, M. (1993). Bootstrap based goodness-of-fit tests. *Metrika* **40** 243-256.
- [11] Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25**. In press.
- [12] Stute, W., Thies, S. and Zhu, L.X. (1996). Nonparametric model checks for regression. Preprint. Univ. of Giessen.



# Bootstrap selection of the smoothing parameter in density estimation under the Koziol-Green model

Jacobo de Uña-Álvarez

*University of Vigo, Spain*

Wenceslao González-Manteiga and Carmen Cadarso-Suárez

*University of Santiago de Compostela, Spain*

*Abstract:* An asymptotic representation of the mean weighted integrated squared error for the kernel estimator of the density under the Koziol-Green model of proportional censorship is obtained for a bootstrap resampling method. A new bandwidth selector based on the bootstrap is consequently proposed. Simulation results for different models using WARPed versions of the estimators show how the bootstrap selector behaves appreciably better than the classical cross-validation method. Finally a real example is analyzed.

*Key words:* Bandwidth selection, bootstrap, Koziol-Green model, mean integrated squared error, warping.

AMS subject classification: 62G07.

## 1 Introduction

A typical situation in survival analysis: let  $Y$  be the variable of interest with continuous distribution function  $F$ , let  $C$  be the right-censoring variable with continuous distribution function  $G$ , and let  $(Z, \delta)$  be the observed pair, i.e.,  $Z = \min(Y, C)$  and  $\delta = 1_{\{Y \leq C\}}$ . The general random censorship model assumes the independence between  $Y$  and  $C$ . Hence  $E(\delta) = \int (1 - G)dF$  and the (continuous) distribution function  $H$  of  $Z$  satisfies  $1 \int H = (1 - F)(1 - G)$ .

The Koziol-Green model of proportional censorship (Koziol and Green,

1976) is an interesting sub-model of the above one obtained by imposing the additional parametric assumption

$$1 - G = (1 - F)^\beta \text{ for some } \beta > 0. \quad (1)$$

A crucial fact about (1) is that the independence between  $Z$  and  $\delta$  characterizes the model (Sethuraman, 1965); this allows to construct hypothesis tests about such a model (see, for example, Herbst, 1992; Henze, 1993).

Under (1) it is true that  $1 - F = (1 - H)^\theta$ , with  $\theta = (1 + \beta)^{-1} = E(\delta)$ . This relation motivates the ACL (Abdushukurov, 1984; Cheng and Lin, 1984) estimator

$$1 - \tilde{F}_n = (1 - H_n)^{\theta_n}, \quad (2)$$

where  $H_n$  is the empirical distribution function of the  $Z$ 's and  $\theta_n$  is the sample mean of the  $\delta$ 's, given the initial sample  $\{(Z_1, \delta_1), \dots, (Z_n, \delta_n)\}$ . The estimator (2) is the maximum-likelihood estimator of the survival function of interest under the model (1).

Here we are interested in the estimation of the density  $f$  of  $Y$  (assumed to exist). A kernel estimator is defined in Section 2. How to choose the bandwidth for it is our main question. Different procedures have been considered and studied in the uncensored case. See Cao *et al.* (1994) for a comparative study. Least squares cross-validation (LSCV) has been adapted to density estimation under proportional censorship by Ghorai and Pattanaik (1993). These authors have established asymptotic optimality, in the sense

$$\frac{ISE_w(h_{CV})}{\inf_{h \in L_n} ISE_w(h)} \rightarrow 1 \text{ a.s.},$$

where the cross-validation bandwidth  $h_{CV}$  is the minimizer of the score function (4) defined in Section 2,  $ISE_w$  is the integrated squared error  $ISE_w(h) = \int (\tilde{f}_h - f)^2 w$ , for a suitable weight function  $w$  (with the role of eliminating endpoint effects),  $\tilde{f}_h$  is the estimator defined in (3), and the set  $L_n$  follows the usual regularity conditions (that can be found in the mentioned work).

In a recent paper González-Manteiga *et al.* (1996) motivate the search for improved methods of bandwidth selection. Although the quality of the resulting density estimator is the crucial question, the rate of convergence for cross-validation type selectors is known to be very slow. In the referred work “smoothed bootstrap” ideas are the base of a new criterion for choosing the bandwidth in censored hazard rate estimation. Here these considerations are translated to the context of density estimation under the Koziol-Green model. In Section 2 we introduce a bootstrap mechanism to select the parameter  $h$  for the estimator  $\tilde{f}_h$ .

We discuss some fast algorithms based on WARP ideas (see Härdle, 1991; Fan and Marron, 1994) in Section 3, introducing methods for computing both the estimator (3) and also the two bandwidth selectors considered here. In Section 4 we present our simulation results for different proportional censorship models. As in González-Manteiga *et al.* (1996), the bootstrap selector behaves appreciably better than the classical cross-validation method. We also present a real example that follows the Koziol-Green model.

## 2 The estimator: LSCV and bootstrap bandwidth selection

A natural kernel estimator for the true parameter of interest  $f$  is given by

$$\tilde{f}_h(y) = \int K_h(y-v)\tilde{F}_n(dv) \simeq \sum_{i=1}^n K_h(y-Z_i)n^{-1}\theta_n(1-H_n(Z_i))^{\theta_n-1} \quad (3)$$

where  $K_h(\cdot) = K(\cdot/h)/h$  is the rescaled kernel function  $K$  with smoothing parameter  $h$  satisfying  $h \rightarrow 0$  and  $nh \rightarrow \infty$  as  $n \rightarrow \infty$ . This estimator was considered by Csörgo and Mielniczuk (1988). These authors proved results on strong consistency, asymptotic normality and Bickel-Rosenblatt type confidence bands for (3).

The LSCV bandwidth  $h_{CV}$  considered here is the minimizer of

$$CV(h) = \int \tilde{f}_h^2 w - 2 \sum_{i=1}^n \tilde{f}_{h,i}(Z_i)n^{-1}\theta_n(1-H_{n,i}(Z_i))^{\theta_n-1}w(Z_i) \quad (4)$$

where  $\tilde{f}_{h,i}$  and  $H_{n,i}$  are the “leave-one-out” versions of  $\tilde{f}_h$  and  $H_n$  respectively, given by

$$\tilde{f}_{h,i}(y) = \sum_{j \neq i} K_h(y-Z_j)(n-1)^{-1}\theta_n(1-H_{n,i}(Z_j))^{\theta_n-1}$$

and

$$H_{n,i}(y) = (n-1)^{-1} \sum_{j \neq i} 1_{\{Z_j \leq y\}}.$$

The function  $CV(h)$  is an estimator of  $MISE_w(h) - \int f^2 w$ , where

$$MISE_w(h) = E \int (\tilde{f}_h - f)^2 w \quad (5)$$

is the mean weighted integrated squared error.

We obtain the next result, giving an asymptotic representation of (5). We make the following assumptions: (i) the density  $f$  is continuous, (ii)



the function  $w$  has compact support contained in  $(0, T)$ , where  $T$  satisfies  $1 - H(T) > 0$ , and (iii) the kernel  $K$  is a square integrable probability density function with compact support.

**Theorem 1** *Under the assumptions (i)-(iii):*

$$MISE_w(h) = AMISE_w(h) + o((nh)^{-1})$$

where

$$AMISE_w(h) = (nh)^{-1} R(K) \int \theta(1 - H)^{\theta-1} fw + \int (K_h * f - f)^2 w$$

and  $R(K) = \int K^2$  (\* denotes convolution).

From now on we define the bootstrap bandwidth selector  $h^*$  as the minimizer of an estimated  $AMISE_w$ . “Smoothed bootstraps” are required to approximate the “bias” part of  $MISE_w$ . We therefore propose the resampling plan:

1. Draw bootstrap resamples  $\{Z_1^*, \dots, Z_n^*\}$  from  $H_n * K_g$ .
2. Generate independent bootstrap resamples  $\{\delta_1^*, \dots, \delta_n^*\}$  from a Bernoulli distribution with parameter  $\theta_n$ .

The distribution  $H_n * K_g$  denotes the one having density  $\hat{h}_g(y) = \int K_g(y - v)H_n(dv)$ . The bootstrap versions of the estimator (3) and the error (5) are respectively

$$\tilde{f}_h^*(y) = \sum_{i=1}^n K_h(y - Z_i^*) n^{-1} \theta_n^* (1 - H_n^*(Z_i^*))^{\theta_n^*-1}$$

and

$$MISE_w^*(h) = E^* \int (\tilde{f}_h^* - \tilde{f}_g)^2 w$$

where  $H_n^*$  is the empirical distribution function of the  $Z^*$ 's and  $\theta_n^*$  is the sample mean of the  $\delta^*$ 's. Similar arguments to those used in the proof of Theorem 1 lead to:

$$AMISE_w^*(h) = (nh)^{-1} R(K) \int \theta_n(1 - H_n * K_g)^{\theta_n-1} \hat{f}_g w + \int (K_h * \hat{f}_g - \tilde{f}_g)^2 w \quad (6)$$

where  $\hat{f}_g = \theta_n(1 - H_n * K_g)^{\theta_n-1} \hat{h}_g$  is another estimator of  $f$  under (1). Then  $h^*$  is the minimizer of (6). This expression shows that our bootstrap design mimics the theoretical  $AMISE_w$  in Theorem 1.

**Remark 1** *We opted to deal with the pilot estimator  $\tilde{f}_g$  instead of the theoretical bootstrap density  $\hat{f}_g$  in the definition of  $MISE_w^*$ . An analogous expression for the asymptotic  $MISE_w^*$  can be obtained using  $\hat{f}_g$ .*

### 3 Warping algorithms

Here we introduce fast algorithms to construct the estimator and to select the bandwidth, both for the LSCV and the bootstrap methods. As in Härdle (1991) and Fan and Marron (1994), the idea is to “bin” the data into an equally spaced grid, so that the number of kernel evaluations can be drastically reduced. Our bins are  $B_z = \left[ \frac{zh}{M}, \frac{(z+1)h}{M} \right)$ ,  $z \in Z$ , where every interval has length  $\delta = h/M$  with  $M \in Z^+$ . We summarize the observed data by the  $n_z = \sum_{i=1}^n 1_{B_z}(Z_i)$ ,  $z \in Z^+$ , the number of observations in the bin  $B_z$ . The large number of kernel differences  $K_h(y - Z_i)$  is approximated by the much smaller set of values  $w_M(k) = K(k/M)$ , for  $k = 1 - M, \dots, M - 1$ , when  $K$  is supported on  $[-1, 1]$ . Fan and Marron (1994) showed that this results in computational speed savings of factors up to 100.

#### 3.1 WARPing the kernel estimator

A WARPing approximation of  $\tilde{f}_h$  at  $B_z$  is given by

$$\tilde{f}_M(z) = (n\delta M)^{-1} \sum_{k=1-M}^{M-1} w_M(k) \theta_n(1 - H_M(z + k))^{\theta_n-1} n_{z+k} \quad (7)$$

where  $1 - H_M(z) = n^{-1} \sum_{k>z} n_k$ . For a fixed  $h$  and letting  $M \rightarrow \infty$ , we have  $\tilde{f}_M(y) \rightarrow \tilde{f}_h(y)$  (the WARPing approximation error decreases with the rounding error  $\delta$ ). This point is illustrated in Figure 1.

#### 3.2 WARPing cross-validation

A WARPed version of the score function (4) in Section 2 involves essentially replacing the conventional kernel estimators by their WARPed versions, although additional work is required to get a rapidly computable version. We define the “leave out bin counts”  $n_z^{-i} = \sum_{j \neq i} 1_{B_z}(Z_j)$ ,  $z \in Z$ , and we define  $1 - H_M^{-i}(z) = (n - 1)^{-1} \sum_{k>z} n_k^{-i}$ .

Now we use the approximations

$$\int \tilde{f}_h^2 w \simeq \delta \sum_z \tilde{f}_M^2(z) w_{M,h}(z)$$

and

$$\begin{aligned} & -2n^{-1} \sum_{i=1}^n \tilde{f}_{h,i}(Z_i) \theta_n(1 - H_{n,i}(Z_i))^{\theta_n-1} w(Z_i) \\ & \simeq -\frac{2}{n-1} \sum_z \tilde{f}_M(z) \theta_n(1 - H_M(z))^{\theta_n-1} n_z w_{M,h}(z) \\ & \quad + \frac{2w_M(0)}{n(n-1)\delta M} \sum_z n_z w_{M,h}(z) (\theta_n(1 - H_M(z))^{\theta_n-1})^2 \end{aligned}$$

where  $w_{M,h}(z)$  denotes the value of  $w$  at the lower limit of  $B_z$ . Our approximated score function becomes

$$\begin{aligned} \overline{CV}(M) = & \delta \sum_z \tilde{f}_M^2(z) w_{M,h}(z) - \frac{2}{n-1} \sum_z \tilde{f}_M(z) \theta_n (1 - H_M(z))^{\theta_n-1} n_z w_{M,h}(z) \\ & + \frac{2w_M(0)}{n(n-1)\delta M} \sum_z n_z w_{M,h}(z) (\theta_n (1 - H_M(z))^{\theta_n-1})^2. \end{aligned} \quad (8)$$

Our proposal is to fix the rounding error  $\delta$  and then to minimize the function (8) in  $M$ .

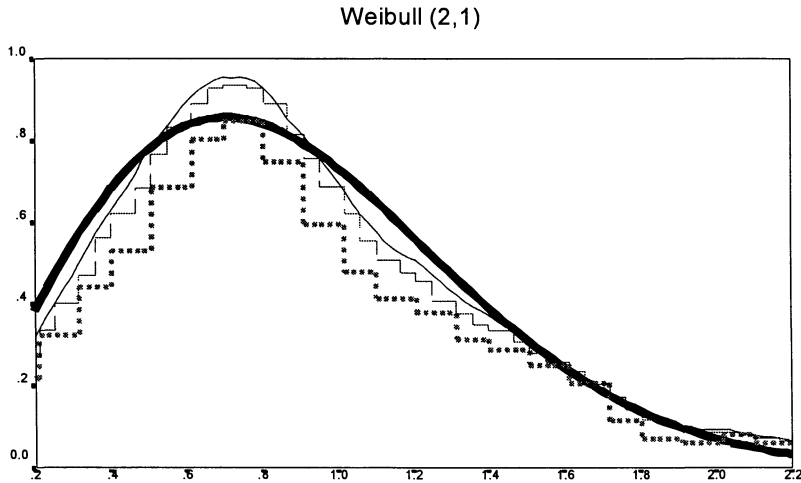


Figure 1: True density function (thick solid line) for a Weibull(2,1) model, the kernel estimator (thin solid line) (computed directly on a sample of 100 uncensored observations of such a model), and two warping approximations of this curve:  $\delta=0.1$  (dotted line) and  $\delta=0.01$  (dashed line). The role of the rounding error becomes clear; the approximation to the kernel estimator gets better as  $\delta$  decreases to zero.

### 3.3 WARPing the smoothed bootstrap

A WARPed version of the function (6) is obtained similarly by replacing the conventional kernel estimates with their corresponding WARPed versions. An approximation of  $AMISE_w^*(h)$  is given by

$$\begin{aligned} \overline{AMISE}^*(M) = & \delta \sum_z (M^{-1} \sum_{k=1-M}^{M-1} w_M(k) \hat{f}_{M_1}(z+k) - \tilde{f}_{M_1}(z))^2 w_{M,h}(z) \\ & + (nh)^{-1} R(K) \delta \sum_z \theta_n (1 - \hat{H}_{M_1}(z))^{\theta_n-1} \hat{f}_{M_1}(z) w_{M,h}(z) \end{aligned} \quad (9)$$

where  $\hat{f}_{M_1}$  and  $\hat{H}_{M_1}$  are the WARPed approximations of  $\hat{f}_g$  and  $H_n * K_g$  with parameter  $M_1 = g/\delta$  at  $B_z$ , given respectively by

$$\hat{H}_{M_1}(z) = (nM_1)^{-1} \sum_{k=1-M}^{M-1} w_{M_1}(k) \bar{n}_{z+k}, \quad \text{with } \bar{n}_z = \sum_{j=-\infty}^z n_j, \text{ and}$$

$$\hat{f}_{M_1}(z) = \theta_n (1 - \hat{H}_{M_1}(z))^{\theta_n - 1} \hat{h}_{M_1}(z), \quad \text{where}$$

$$\hat{h}_{M_1}(z) = (n\delta M)^{-1} \sum_{k=1-M}^{M-1} w_{M_1}(k) n_{z+k}.$$

## 4 Simulation study. Example

### 4.1 Simulation study

In this subsection we compare LSCV and bootstrap bandwidth selectors for moderate sample sizes. We considered three underlying densities:

**Weibull**  $(\alpha, \lambda)$ . The density of interest is taken as  $f = f_{\alpha, \lambda}$ , satisfying  $f(x) = f_{\alpha, \lambda}(x) = \alpha \lambda (\lambda x)^{\alpha-1} e^{-(\lambda x)^\alpha} 1_{[0, \infty)}(x)$  with  $\alpha, \lambda > 0$ .

**Gumbel**  $(\alpha, \lambda)$ . Its density is  $f(x) = f_{\alpha, \lambda}(x) = \alpha \lambda e^{\lambda x - \alpha(e^{\lambda x} - 1)} 1_{[0, \infty)}(x)$  with  $\alpha, \lambda > 0$ .

**Truncated normal model.** We consider the density of the random variable  $Y = X \mid X \geq 0$ , where  $X \in N(\mu, \sigma)$ ; that is,  $f(x) = \frac{\phi_{\mu, \sigma}(x)}{1 - \Phi_{\mu, \sigma}(0)} 1_{[0, \infty)}(x)$ , where  $\phi_{\mu, \sigma}(x) = \Phi'_{\mu, \sigma}(x)$  and  $\Phi_{\mu, \sigma}$  is the distribution of  $X$ .

This simulation study was carried out as in González-Manteiga *et al.* (1996) (see this article for details).

Table 1 contains the results of 1.000 trials of sample size 100 corresponding to the following models:

- Weibull models with parameters  $\lambda=1$  and  $\alpha=1, 2, 3$  without censoring (denoted by W(1,1), W(2,1) and W(3,1)) and also with 25% of censoring (denoted by CW(1,1), CW(2,1) and CW(3,1)).
- Gumbel models with parameters  $\lambda=1$  and  $\alpha=1, 2, 3$  without censoring (G(1,1), G(2,1) and G(3,1)) and with 25% of censoring (CG(1,1), CG(2,1) and CG(3,1)).
- Truncated normal distributions with parameters  $\mu=1$  and  $\sigma=0.5$  for an uncensored situation and also with a censoring of 25% (N(1,0.5), CN(1, 0.5)).

Although the only sample size in the simulations presented here is  $n=100$ , similar results were observed for  $n=50$  and  $n=200$ . The triangular kernel was used.

| MODEL     | Mean  |       | Std.Dev. |       |
|-----------|-------|-------|----------|-------|
|           | hcv   | h*    | hcv      | h*    |
| W(1,1)    | 20.96 | 15.21 | 14.54    | 8.50  |
| CW(1,1)   | 21.57 | 18.27 | 13.55    | 9.69  |
| W(2,1)    | 16.64 | 10.75 | 19.14    | 9.21  |
| CW(2,1)   | 16.01 | 11.38 | 20.19    | 8.61  |
| W(3,1)    | 21.62 | 14.39 | 26.47    | 12.32 |
| CW(3,1)   | 21.45 | 17.31 | 26.64    | 13.75 |
| G(1,1)    | 22.84 | 15.78 | 22.97    | 11.15 |
| CG(1,1)   | 23.46 | 16.36 | 22.44    | 9.72  |
| G(2,1)    | 42.58 | 30.63 | 33.01    | 19.02 |
| CG(2,1)   | 44.12 | 34.59 | 31.60    | 19.39 |
| G(3,1)    | 60.83 | 45.94 | 42.03    | 27.14 |
| CG(3,1)   | 64.29 | 53.52 | 42.63    | 29.25 |
| N(1,0.5)  | 14.82 | 10.41 | 16.78    | 8.81  |
| CN(1,0.5) | 19.16 | 13.17 | 25.12    | 10.38 |

Table 1: Mean and standard deviation of the integrated squared error (L<sub>2</sub>-norm) of the kernel density estimator with  $h_{CV}$  and  $h^*$  bandwidths along 1.000 trials of size 100.

The values in Table 1 are the mean and the standard deviation of the  $ISE_w$  (L<sub>2</sub>-norm)  $\int (\tilde{f}_{\hat{h}} - f)^2 w$  along the 1.000 samples of size 100. The columns headed  $h_{CV}$  report the  $ISE_w$  for estimates based on cross-validation bandwidths (resulting from the minimization of  $\overline{CV}(M)$  in expression (8)) and the columns headed  $h^*$  report the  $ISE_w$  for estimates based on bootstrap bandwidths (resulting from the minimization of  $\overline{AMISE}^*(M)$  in expression (9)). Both minimizers were taken as the global minimizer over a fine grid. In both cases the rounding error in the WARP approximation was  $\delta=0.01$  and the weighting function was  $w(u) = 1_{[H^{-1}(0.05), H^{-1}(0.95)]}(u)$ . Finally, the pilot bandwidth  $g$  used is given by  $g = \left[ \int (K'')^2 \mu_2(K)^{-1} \hat{a}^{-1} n^{-1} \right]^{1/7}$ , where  $\hat{a}$  is the estimator of  $\int (h_Z^{(3)})^2 (h_Z$  the density of  $Z$ ) and  $\mu_2(K) = \int t^2 K(t) dt$ . For a sample from a normal density, the pilot bandwidth  $g$  is asymptotically optimal for the purpose of estimating the density curvature by the curvature of the kernel estimator (for details see Cao *et al.*, 1994).

| MODEL     | Mean   |        | Std. Dev. |       |
|-----------|--------|--------|-----------|-------|
|           | hcv    | h*     | hcv       | h*    |
| W(1,1)    | 146.55 | 128.70 | 49.91     | 38.19 |
| CW(1,1)   | 135.03 | 117.97 | 45.48     | 33.95 |
| W(2,1)    | 105.86 | 89.47  | 51.91     | 38.91 |
| CW(2,1)   | 96.92  | 89.27  | 50.09     | 35.47 |
| W(3,1)    | 102.39 | 87.54  | 50.99     | 38.23 |
| CW(3,1)   | 96.80  | 93.93  | 48.93     | 39.09 |
| G(1,1)    | 113.94 | 101.83 | 50.01     | 38.84 |
| CG(1,1)   | 107.99 | 91.20  | 47.25     | 30.21 |
| G(2,1)    | 124.45 | 112.14 | 47.86     | 37.88 |
| CG(2,1)   | 117.35 | 101.75 | 44.32     | 31.47 |
| G(3,1)    | 128.52 | 117.05 | 45.89     | 37.47 |
| CG(3,1)   | 120.48 | 106.84 | 42.83     | 32.25 |
| N(1,0.5)  | 102.81 | 89.58  | 49.98     | 38.96 |
| CN(1,0.5) | 91.17  | 82.61  | 48.10     | 33.43 |

Table 2: Mean and standard deviation of the integrated absolute error (L<sub>1</sub>-norm) of the kernel density estimator with  $h_{CV}$  and  $h^*$  bandwidths along 1.000 trials of size 100.

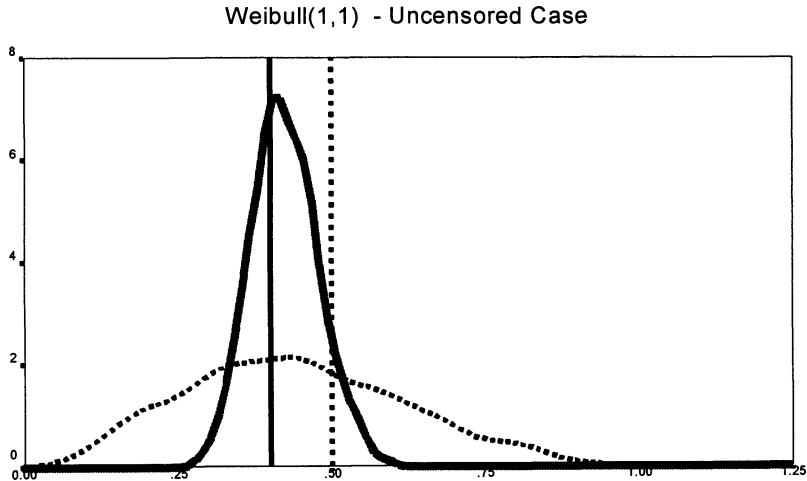


Figure 2: Kernel estimator of the densities of  $h_{CV}$  (dashed line) and  $h^*$  (solid line) using Gaussian kernel and smooth cross-validation bandwidth, based on 1.000 trials of size 100 of a Weibull(1,1) model (uncensored case). The vertical lines represent the values of  $h_{l2}$  (solid line) and  $h_{l1}$  (dotted line).

Table 2 is devoted to the L<sub>1</sub>-norm  $\int |\tilde{f}_h - f| w$ . Tables 1 and 2 show that the L<sub>1</sub> and L<sub>2</sub> norms are more concentrated around their means for the bootstrap selector than for the cross-validation bandwidth. The L<sub>1</sub> and

$L_2$  norms tend to be smaller with the bootstrap bandwidth than that with cross-validation. To show the accuracy of both bandwidth selectors figures 2 (uncensored case) and 3 (censored) present the kernel estimators of the densities of the two selectors  $h_{CV}$  and  $h^*$ , using Gaussian kernel and smooth cross-validation bandwidth, based on a sample of 1.000 different values from the models  $W(1,1)$  and  $CW(1,1)$ , as well as some approximations of the  $h_{L_2,w}$  and  $h_{L_1,w}$  bandwidths. These bandwidths were computed by minimizing the Monte Carlo approximation, based on 1.000 trials of the  $MISE_w, E \int (\tilde{f}_h - f)^2 w$ , and the  $MIAE_w, E \int |\tilde{f}_h - f| w$ , over a grid of  $h$  values, ranging from 0.1 to 1.1 with a step of 0.05. We conclude that the performance of  $h^*$  is far superior to that of  $h_{CV}$ . This is what we expected, because as noted in Cao *et al.* (1994) at least in the uncensored case we have

$$\frac{h_{CV} - h_f}{h_f} = O_P(n^{-1/10}) \quad \frac{h^* - h_f}{h_f} = O_P(n^{-5/14})$$

where  $h_f$  denotes the optimum bandwidth which minimizes  $MISE$ .

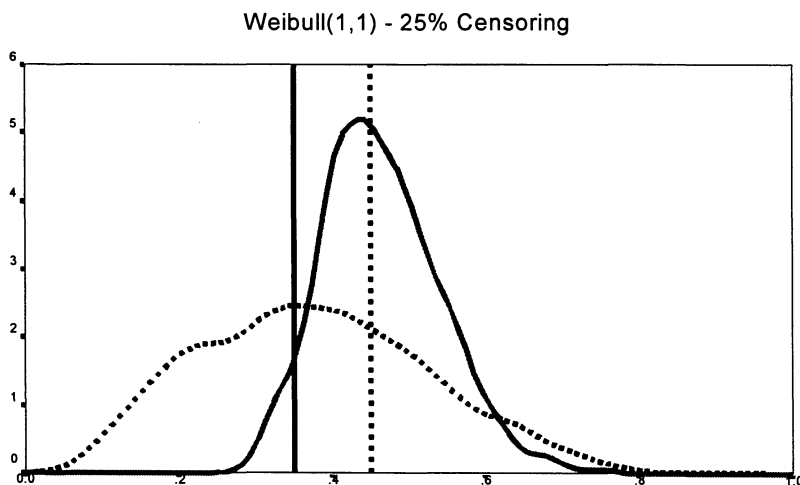


Figure 3: Kernel estimator of the densities of  $h_{CV}$  (dashed line) and  $h^*$  (solid line) using Gaussian kernel and smooth cross-validation bandwidth, based on 1.000 trials of size 100 of a Weibull(1,1) model (25% censoring). The vertical lines represent the values of  $h_{L2}$  (solid line) and  $h_{L1}$  (dotted line).

To illustrate the computational cost of the WARPing approximation when used in the algorithms for finding  $h^*$  and  $h_{CV}$ , the CPU times (relative to the minimum of all of them) are summarized in Table 3. While CPU

time increases slowly with the sample size,  $n$ , a sharp increase occurs when the rounding error,  $\delta$ , gets small. We suggest the choice  $\delta=0.01$  in practice. This seems to give a good approximation of the kernel estimator (see Figure 1) at a reasonable computational cost. (See Fan and Marron, 1994) for a comparative study of the CPU time of the WARPing approximation and the kernel estimator).

| Sample Size | $\delta=0.05$ |      | $\delta=0.01$ |       | $\delta=0.005$ |        |
|-------------|---------------|------|---------------|-------|----------------|--------|
|             | hCV           | h*   | hCV           | h*    | hCV            | h*     |
| n=50        | 1.33          | 1.00 | 23.78         | 14.08 | 97.82          | 52.69  |
| n=100       | 5.46          | 1.17 | 91.88         | 21.12 | 373.83         | 81.70  |
| n=500       | 5.65          | 1.35 | 107.68        | 20.54 | 433.19         | 107.95 |

Table 3: Relative CPU times of the computations of LSCV and bootstrap bandwidths for samples of sizes  $n=50$ , 100 and 500 and rounding errors  $\delta=0.05$ , 0.01 and 0.005 in the WARPing approximation.

## 4.2 Example: PCB-Liver data

Between January, 1974, and May, 1984, the Mayo Clinic conducted a double-blinded randomized trial in Primary Biliary Cirrhosis (PCB) of the liver. A total of  $n=312$  patients agreed to participate in this clinical trial. The data were analyzed in 1986 for presentation in clinical literature (see Fleming *et al.*, 1991). By July, 1986, 125 of the 312 patients had died, resulting in a high proportion of censoring data (60%).

|           | Sample Size | Deaths | Censored (%) | OR (95% CI)                                      | Median | Mean  |
|-----------|-------------|--------|--------------|--------------------------------------------------|--------|-------|
| GROUP I   | 163         | 32     | 131 (80.37%) | 1 Reference                                      | *      | 10.41 |
| GROUP II  | 63          | 31     | 32 (50.79%)  | 3.5 (2.2,5.8)                                    | 7.66   | 7.22  |
| GROUP III | 50          | 31     | 19 (38.00%)  | 6.45 (3.9,10.7)                                  | 4.63   | 5.47  |
| GROUP IV  | 36          | 31     | 5 (13.89%)   | 18.9 (11.0,32.5)                                 | 2.35   | 2.60  |
| TOTAL     | 312         | 125    | 187 (59.44%) | * More than 50% people alive at the closing date |        |       |

Table 4: Sample size, number of deaths, percentage of censoring, Odds-Ratio and estimated median and mean for the four groups of patients considered in the PCB-Liver data example according to their prognosis bilirubin levels ( $<1.45$ ,  $1.45-3.25$ ,  $3.25-6.75$  and  $>6.75$ ).

One of the most important risk factor for the survivorship of PCB is the serum bilirubin level (Fleming *et al.*, 1991). As in the referred work, we shall consider four groups of patients, according to their prognosis bilirubin levels: Group I (bilirubin $<1.45$ ), Group II ( $1.45-3.25$ ), Group III ( $3.25-6.75$ ) and Group IV ( $>6.75$ ). Descriptive statistics appear in Table 4, revealing that the survival time decreases as albumin level increases. Whereas



Group I presents an 80.3% of censored data, Group II has an amount around 50.6%, Group III 38% and Group IV, only 14%. The four groups were tested to follow the Koziol-Green model using the test proposed by Henze (1993). The approximated  $p$ -values were 0.1627, 0.1552, 0.1765 and 0.46764, respectively, failing to reject the proportional censoring model for each group. The estimator (3) of each density function can therefore be safely used. In all cases, a gaussian kernel was chosen. For each group, the weighting function considered here was  $w(u) = 1_{[a,b]}(u)$  where  $a$  and  $b$  are, respectively, the 5% and 95% percentiles of the corresponding observed survival time. The selected bootstrap bandwidth minimizing their respective function  $\overline{AMISE}^*(M)$  (with rounding error  $\delta=0.1$ ) were  $h_I^*=3.9$ ,  $h_{II}^*=4.01$ ,  $h_{III}^*=3.3$  and  $h_{IV}^*=1.73$ . The density functions estimates are plotted together in Figure 4. Looking at this figure, one observes that the density estimates reflect the behaviour of the survival time, revealing a great amount of probability around the median values, in agreement with results presented above.

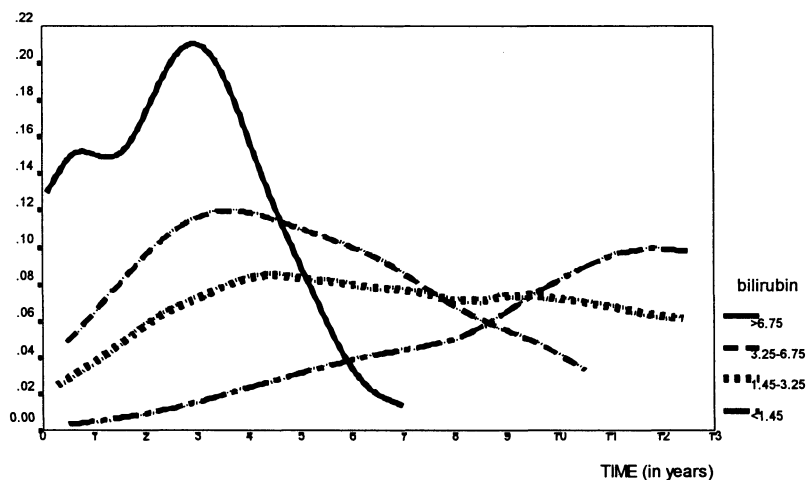


Figure 4: Kernel density estimators for the groups in the PCB-Liver data example: bilirubin  $> 6.75$  (solid line),  $3.25-6.75$  (dashed line),  $1.45-3.25$  (dotted line) and  $< 1.45$  (dashed-dotted line). The figure supports the numerical results in Table 4.

## 5 Conclusions

We proposed a smoothed bootstrap selector of the smoothing parameter in kernel density estimation when the Koziol-Green model of proportional

ensorship holds. We presented a mathematical analysis supported by simulation. It turned out that our proposal behaves convincingly better than the cross-validation selector. We gave fast implementation using WARP-ing methods. We showed the practical interest of the introduced techniques through the analysis of real medical data sets.

## Acknowledgements

The work of the second and the third author was partially supported by the project XUGA20701B96 of the Xunta de Galicia and by the DGES grant PB95-0826.

## References

- [1] Abdushukurov, A. A. (1984). On some estimates of the distribution function under random censorship. Conference of Young Scientists, Math. Inst. Acad. Sci. Uzbek SSR, Tashkent. VINITI n° 8756-V (in Russian).
- [2] Cao, R., Cuevas, A. and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Comput. Statist. Data Anal.* **17** 153-176.
- [3] Cheng, P. E. and Lin, G.D. (1984). Maximum likelihood estimation of a survival function under the Koziol-Green proportional hazard model. Technical Report, Institute of Statistics, Academia Sinica.
- [4] Csörgo, S. and Mielniczuk, J. (1988). Density estimation in the simple proportional hazards model. *Statistics and Probability Letters* **6** 419-426.
- [5] Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics* **3** 35-56.
- [6] Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes & Survival Analysis*. New York: Wiley-Interscience.
- [7] Ghorai, J. K. and Pattanaik, L. M. (1993). Asymptotically optimal bandwidth selection of the kernel density estimator under the proportional hazards model. *Commun. Statist. Theory and Methods* **22** 1383-1401.
- [8] González-Manteiga, W., Cao, R. and Marron, J. S. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *J.Am. Statist.Assoc.* **91** 1130-1140.
- [9] Härdle, W. (1991). *Smoothing Techniques with Implementation in S*. Springer-Verlag.
- [10] Henze, N. (1993). A quick omnibus test for the proportional hazards

- model of random censorship. *Statistics* **24** 253-263.
- [11] Herbst, T. (1992). Test of fit with the Koziol-Green model for random censorship. *Statistics & Decisions* **10** 163-171.
  - [12] Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises statistic for randomly censored data. *Biometrika* **63** 465-474.
  - [13] Sethuraman, J. (1965). On a characterization of the three limiting types of the extreme. *Sankhyā A* **27**, 357-364.

# On multivariate rank regression

**Biman Chakraborty**

*Indian Statistical Institute, Calcutta, India*

**Probal Chaudhuri**

*Indian Statistical Institute, Calcutta, India*

*Abstract:* An extension of rank regression techniques to multivariate linear models is proposed and studied. Unlike the co-ordinatewise rank regression techniques considered by some earlier authors, our approach is affine equivariant, and it is based on a transformation and retransformation procedure originally developed by Chakraborty and Chaudhuri (1996, 1997) for constructing an affine equivariant version of multivariate median. Affine equivariance is expected to lead to superior statistical performance of our procedure compared to other non-equivariant procedures especially in the presence of substantial correlations among different response variables in multi-response problems. Some of the statistical properties of the proposed multivariate rank regression estimates are discussed, and a few results based on numerical investigation of the performance of these estimates are presented.

*Key words:* Affine equivariance, Hodges-Lehmann estimate, multivariate linear model, statistical efficiency, transformation retransformation estimates, Wilcoxon's rank scores.

AMS subject classification: Primary 62H12, 62J05; secondary 62F35.

## 1 Introduction: multivariate linear model and rank regression

Linear model is a widely used statistical tool for empirical analysis to understand and make inference about the nature of inter-dependence that exists among different variables in the data. Perhaps it will not be an overstatement to say that various forms of linear model pervade almost every

area of applied research where statistics has its scope of being effectively used. Here our focus will be on multivariate linear models of the form

$$\mathbf{Y} = \mathbf{\Gamma}\mathbf{X} + \mathbf{e} \quad , \quad (1)$$

where  $\mathbf{Y}$  is a  $d$ -dimensional column vector of dependent or response variables,  $\mathbf{X}$  is a  $p$ -dimensional column vector of independent or regressor variables, and  $\mathbf{\Gamma}$  is the  $d \times p$  matrix of unknown coefficient parameters that determine how different regressor variables jointly influence different response variables. This matrix is to be estimated from the observed data  $(\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, \mathbf{Y}_2), \dots, (\mathbf{X}_n, \mathbf{Y}_n)$ , and the term  $\mathbf{e}$  in (1) is a  $d$ -dimensional column vector of random errors representing the random deviation of a data point from the linear model. In the special case when the response is real valued (i.e. when  $d = 1$ ), rank regression techniques have been proposed and extensively studied as an alternative to traditional least squares regression by several statisticians [see e.g. Lehmann (1963a, 1963b, 1964), Adichi (1967, 1978), Koul (1969, 1970), Puri and Sen (1969, 1973), Jureckova (1971, 1973), Jaeckel (1972), Hettmansperger and McKean (1977, 1978, 1983)]. These authors explored various extensions of rank based methods, which were originally developed for nonparametric inference in one and two sample univariate location problems, into very general linear models including standard ANOVA models. A primary motivation behind considering rank regression is the lack of robustness in least squares regression, which is known to have very poor performance when the random error  $\mathbf{e}$  in the linear model (1) happens to follow non-Gaussian distributions especially those with heavy tails. Higher statistical efficiency of rank based nonparametric procedures compared to the inference based on sample means in one and two sample location problems involving univariate non-Gaussian data is known to extend for parameter estimates and related inference based on rank regression in linear models with univariate response. An excellent review of various rank based statistical methods for linear models with real valued response can be found in Draper (1988) and in fascinating comments and discussion that Draper's expository article was successful in generating from leading experts in robust regression in linear models.

Unfortunately most of the work documented in the existing literature on rank regression is restricted to univariate response. While many practical situations (e.g. when a medical scientist is interested in studying the relationship between the age of an individual and his/her systolic and diastolic blood pressures) do involve multi-response problems, virtually very little is available in the literature other than least squares techniques when the response  $\mathbf{Y}$  in the linear model (1) happens to be multi-dimensional

in nature (i.e.  $d > 1$ ). Rao (1988) and Koenker and Portnoy (1990) considered robust estimation of parameters in multi-response linear regression problems and suggested the use of univariate least absolute deviations method for each co-ordinate of the response vector. Davis and McKean (1993) have extensively studied the co-ordinatewise extension of rank regression from univariate to multivariate response problems. These authors have derived some interesting statistical properties of their proposed estimates and tests and reported some results on numerical performance of the procedures. One serious drawback of co-ordinatewise extension of rank regression as well as that of least absolute deviations regression is that such extensions fail to take into account the inter-dependence that exists among the real valued components of the response vector, and in practice it may not be appropriate to ignore the correlation present among different response variables. Such an approach for multivariate linear models leads to parameter estimates that are not equivariant and to statistical tests that are not invariant under general affine transformation of the data, and it is known that procedures that lack affine equivariance/invariance may have very poor statistical performance in the presence of substantial correlation among the components of the response vector. This issue has been discussed and investigated by Bickel (1964), Brown and Hettmansperger (1987, 1989) and Chakraborty and Chaudhuri (1996, 1997) in the context of multivariate location problems. It will be appropriate to note here that Bai, Chen, Miao and Rao (1990) proposed to estimate  $\Gamma$  in the linear model (1) by minimizing the sum  $\sum_{i=1}^n \|\mathbf{Y}_i - \Gamma \mathbf{X}_i\|$  w.r.t.  $\Gamma$ , where for a  $d$ -dimensional vector  $\mathbf{x} = (x_1, x_2, \dots, x_d)$ ,  $\|\mathbf{x}\|$  = the usual Euclidean norm of  $\mathbf{x} = (x_1^2 + x_2^2 + \dots + x_d^2)^{1/2}$ , and such an estimate of  $\Gamma$  can be viewed as a generalization of the notion of spatial median [see e.g. Haldane (1948), Brown (1983)] in linear models. While this leads to estimates that are equivariant under rotation or orthogonal transformation of the response vector, such estimates still lack equivariance under general affine transformation of the response. It is known that for multivariate data with correlated variables, spatial median may have poor statistical efficiency compared to affine invariant sample mean vector [see Brown (1983), Chaudhuri (1992a), Chakraborty, Chaudhuri and Oja (1997)]. Further, the lack of scale equivariance makes spatial median as well as its generalization in linear models practically useless when different real valued components of the response vector  $\mathbf{Y}$  are measured in different scales or when the response variables have different degrees of statistical variation.

Chakraborty (1996) proposed and studied in detail an affine equivariant extension of least absolute deviations regression in multi-response linear model problems using a data driven transformation and retransformation

approach, which was used earlier by Chakraborty and Chaudhuri (1996, 1997) to construct an affine equivariant version of multivariate median. Such a transformation and retransformation technique converts nonequivariant estimates into equivariant ones and thereby improves upon the performance of the estimates as measured by appropriate statistical efficiency criteria. Our goal in this paper is to use the same transformation and retransformation strategy for developing affine equivariant rank regression techniques that can be used in the analysis of data following multivariate linear models. Chakraborty (1996) gave a convenient algorithm (called TREMMER) for computing the estimate of the parameter matrix  $\mathbf{\Gamma}$  in (1) and amply demonstrated how resampling strategies like the bootstrap can be invoked to estimate finite sample-variance covariance matrix of such a parameter estimate. In Section 2 that follows, we will describe how one can appropriately modify TREMMER to come up with affine equivariant rank regression procedures for multi-response linear models. Such a modification inherits the nice statistical properties of TREMMER, and in the special case of regression based on Wilcoxon's rank scores or equivalently the linear regression analogs of Hodges-Lehmann type estimates [see e.g. Chaudhuri (1992b)], this modification takes a simplified and elegant form that makes the implementation of the methodology as well as investigation into its statistical properties quite convenient. In Section 3, we will present some results based on numerical studies that were undertaken to investigate the performance of the proposed methodology. We will discuss results from small sample simulation experiments that yield strong evidence for superior performance of transformation retransformation rank regression estimates in multi-response linear model problems when compared with traditional least squares and co-ordinatewise least absolute deviations estimates if the residuals in the linear model have elliptic non-Gaussian distributions with heavy tails. We will also report analysis of two real data sets each with bivariate response in an attempt to demonstrate the implementation of the methodology in real data and how it outperforms some competing non-equivariant procedures. Section 4 will conclude the paper with some remarks on the issues that have become transparent in course of our present research, and there we will try to discuss briefly some of the open problems that require further research.

## 2 The transformation and retransformation procedure

Let us now focus our attention on the data points  $(\mathbf{X}_i, \mathbf{Y}_i)$ 's, which are assumed to satisfy the linear model (1). Suppose that  $n > d + p$ , and  $\alpha$  is

a subset of size  $d + p$  of the set of indices  $\{1, 2, \dots, n\}$ . Following the notation used by Chakraborty (1996), we will write  $\alpha = \{i_1, \dots, i_p, j_1, \dots, j_d\}$  and denote by  $\mathbf{W}(\alpha)$  the  $p \times p$  matrix whose columns are the vectors  $\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_p}$  and by  $\mathbf{Z}(\alpha)$  the  $d \times p$  matrix whose columns are the vectors  $\mathbf{Y}_{i_1}, \dots, \mathbf{Y}_{i_p}$ . We will assume that  $\mathbf{W}(\alpha)$  is invertible and form the  $d \times d$  matrix  $\mathbf{E}(\alpha)$  that consists of the columns  $\mathbf{Y}_{j_1} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}\mathbf{X}_{j_1}, \dots, \mathbf{Y}_{j_d} - \mathbf{Z}(\alpha)\{\mathbf{W}(\alpha)\}^{-1}\mathbf{X}_{j_d}$ . The matrix  $\mathbf{E}(\alpha)$  too is assumed to be non-singular, and we define the transformed response vectors as  $\mathbf{Z}_l^{(\alpha)} = \{\mathbf{E}(\alpha)\}^{-1}\mathbf{Y}_l$  for  $1 \leq l \leq n$  and  $l \notin \alpha$ . Suppose now that we perform rank regression on each co-ordinate of  $\mathbf{Z}_l^{(\alpha)}$  separately with  $\mathbf{X}_l$  as the regressor as has been done in Davis and McKean (1993), and the resulting estimate of the matrix of coefficient parameters is denoted by  $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$ . In other words,  $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$  is obtained by minimizing (w.r.t.  $\mathbf{\Lambda}$ ) a dispersion function  $\mathcal{D}(\mathbf{\Lambda})$  (say), which is a simple multivariate extension of Jaeckel's dispersion function [see Jaeckel (1972)] based on residuals and their ranks computed from a linear model. In this case  $\mathcal{D}(\mathbf{\Lambda})$  is a function of the real valued co-ordinates of the multivariate residuals  $\mathbf{Z}_l^{(\alpha)} - \mathbf{\Lambda}\mathbf{X}_l$  with  $1 \leq l \leq n$ ,  $l \notin \alpha$  and their ranks [see Davis and McKean (1993)]. Finally, the transformation retransformation estimate of  $\mathbf{\Gamma}$  is obtained by retransforming  $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$  by the matrix  $\mathbf{E}(\alpha)$  as follows

$$\hat{\mathbf{\Gamma}}_n^{(\alpha)} = \mathbf{E}(\alpha)\hat{\mathbf{\Lambda}}_n^{(\alpha)} . \quad (2)$$

In view of the definition of  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ , we now have the following result, which asserts that it is an affine equivariant estimate of  $\mathbf{\Gamma}$ . As a matter of fact, this result is the analog of Chakraborty's (1996) Theorem 2.1 in the context of rank regression.

**Result 1** *Suppose that  $\mathbf{A}$  is a fixed  $d \times d$  non-singular matrix. Then the transformation retransformation estimate computed from the transformed data points  $(\mathbf{X}_1, \mathbf{A}\mathbf{Y}_1), (\mathbf{X}_2, \mathbf{A}\mathbf{Y}_2), \dots, (\mathbf{X}_n, \mathbf{A}\mathbf{Y}_n)$  in the same way as above (i.e. using the same index set  $\alpha$ ) will be  $\mathbf{A}\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ . Further, if the response vector  $\mathbf{Y}_i$  is transformed to  $\mathbf{Y}_i - \mathbf{G}\mathbf{X}_i$  for each  $i = 1, 2, \dots, n$ , where  $\mathbf{G}$  is a fixed  $d \times p$  matrix, the transformation retransformation estimate gets transformed to  $\hat{\mathbf{\Gamma}}_n^{(\alpha)} - \mathbf{G}$ .*

**Proof:** In view of the construction of  $\mathbf{Z}(\alpha)$ , when the  $\mathbf{Y}_i$ 's are transformed to  $\mathbf{A}\mathbf{Y}_i$ 's,  $\mathbf{Z}(\alpha)$  gets transformed to  $\mathbf{A}\mathbf{Z}(\alpha)$ , and consequently the transformation matrix  $\mathbf{E}(\alpha)$  is transformed to  $\mathbf{A}\mathbf{E}(\alpha)$ . On the other hand, since the  $\mathbf{Z}_l^{(\alpha)}$ 's remain invariant under non-singular linear transformation of the  $\mathbf{Y}_i$ 's, so does the estimate  $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$ , which is obtained by performing co-ordinatewise rank regression of the  $\mathbf{Z}_l^{(\alpha)}$ 's on the regressor vectors  $\mathbf{X}_l$ 's.



Hence, the transformation retransformation estimate will be transformed to  $\mathbf{A}\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  as it has been claimed in the statement of the result. Next note that when each of the  $\mathbf{Y}_i$ 's is transformed as  $\mathbf{Y}_i - \mathbf{G}\mathbf{X}_i$ , the matrix  $\mathbf{Z}(\alpha)$  becomes  $\mathbf{Z}(\alpha) - \mathbf{G}\mathbf{W}(\alpha)$ . However, the matrix  $\mathbf{E}(\alpha)$  remains unaltered after such a transformation of the response. Since each of the  $\mathbf{Z}_l^{(\alpha)}$ 's gets transformed as  $\mathbf{Z}_l^{(\alpha)} - \{\mathbf{E}(\alpha)\}^{-1}\mathbf{G}\mathbf{X}_l$ 's in this case, the equivariance of co-ordinatewise rank regression estimate implies that  $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$  will now become  $\mathbf{\Lambda}_n^{(\alpha)} - \{\mathbf{E}(\alpha)\}^{-1}\mathbf{G}$ . The proof is now complete in view of the way  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  has been formed by retransforming  $\hat{\mathbf{\Lambda}}_n^{(\alpha)}$  into  $\mathbf{E}(\alpha)\hat{\mathbf{\Lambda}}_n^{(\alpha)}$ .  $\square$

## 2.1 Selection of the optimal data driven transformation

Since the estimate  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  depends on the choice of the transformation matrix  $\mathbf{E}(\alpha)$ , a question that naturally arises at this point is how to choose the subset of indices  $\alpha$ . This question has been dealt with in other situations by Chakraborty and Chaudhuri (1996, 1997), Chakraborty (1996) and Chakraborty, Chaudhuri and Oja (1997), who used transformation and retransformation techniques in different multivariate estimation problems. Depending on the nature of the problem, these authors have determined the form of the optimal transformation  $\mathbf{E}(\alpha)$  and suggested appropriate data driven selection procedure for the optimal subset of indices  $\alpha$ . All these procedures for choosing the optimal transformation matrix, however, are based on the common idea of minimizing the generalized variance (i.e. the determinant of the dispersion matrix) of the multivariate location or regression estimate. The motivation for looking at the generalized variance comes from the fact that it is proportional to the volume of the concentration ellipsoid associated with the sampling distribution of the estimate which is usually normal for large samples. We will now state a result that asserts that under suitable regularity conditions  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  is a  $n^{1/2}$ -consistent and asymptotically normal estimate of the parameter matrix  $\mathbf{\Gamma}$  in the linear model (1).

**Result 2** *Fix an  $\alpha$ . Suppose that the distribution of the  $(\mathbf{X}_i, \mathbf{Y}_i)$ 's and the nature of the dispersion function  $\mathcal{D}(\mathbf{\Lambda})$  are such that  $n^{1/2}$ -consistency and asymptotic normality of the co-ordinatewise rank regression estimates holds. For example the regularity conditions used in Davis and McKean (1993), who considered co-ordinatewise rank regression will be sufficient for this purpose. Then conditioned on  $\alpha$  and the  $(\mathbf{X}_i, \mathbf{Y}_i)$ 's with  $i \in \alpha$ , the asymptotic distribution of  $n^{1/2}(\hat{\mathbf{\Gamma}}_n^{(\alpha)} - \mathbf{\Gamma})$  is multivariate normal with zero mean and a variance covariance matrix that depends on the transformation matrix  $\mathbf{E}(\alpha)$ .*

**Proof:** Let us fix an  $\alpha$  and argue conditionally given the  $(\mathbf{X}_i, \mathbf{Y}_i)$ 's with  $i \in \alpha$ . Note that since  $\hat{\Lambda}_n^{(\alpha)}$  is obtained by performing co-ordinatewise rank regression of the transformed response vectors  $\mathbf{Z}_i^{(\alpha)}$ 's on the covariates  $\mathbf{X}_i$ 's, it will be a  $n^{1/2}$ -consistent and asymptotically normal estimate of  $\{\mathbf{E}(\alpha)\}^{-1}\mathbf{\Gamma}$  under appropriate regularity conditions as assumed in the statement of the result. The proof is now complete if we recall that  $\hat{\mathbf{\Gamma}}_n^{(\alpha)} = \mathbf{E}(\alpha)\hat{\Lambda}_n^{(\alpha)}$  and use the fact that linear transformation preserves multivariate normality as well as  $n^{1/2}$ -consistency.  $\square$

However, the conditional asymptotic dispersion matrix of  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  depends on  $\mathbf{E}(\alpha)$  in a rather complex way, and it is hardly useful in providing any insight regarding the optimal choice of  $\alpha$  in a general situation. Alternatively, one can try to use resampling techniques (e.g. the bootstrap) to estimate the sampling variation in  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$ , and then select an optimal  $\mathbf{E}(\alpha)$  based on this estimate. However, it does not seem to be a feasible approach in practice in view of the enormous amount of computation that any form of resampling estimation of the dispersion of  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  will require for different choices of  $\alpha$ .

Suppose now that  $\mathbf{e}$  has an elliptically symmetric distribution with a density of the form  $\{\det(\Sigma)\}^{-1/2}f(\mathbf{e}^T\Sigma^{-1}\mathbf{e})$ , where  $\Sigma$  is a  $d \times d$  positive definite matrix, and  $f$  is a probability density function on the real line. Let us write  $\{\Sigma^{-1/2}\mathbf{E}(\alpha)\}^{-1} = \mathbf{R}(\alpha)\mathbf{J}(\alpha)$ , where  $\mathbf{R}(\alpha)$  is a diagonal matrix with positive diagonal entries, and  $\mathbf{J}(\alpha)$  is a matrix whose rows are of unit length, and define  $\mathbf{D}(\alpha)$  to be the symmetric  $d \times d$  matrix whose  $(i, j)$ -th element is  $\sin^{-1} \gamma_{ij}$ ,  $\gamma_{ij}$  being the Euclidean inner product of the  $i$ -th and the  $j$ -th row of  $\mathbf{J}(\alpha)$ . Then it was proved by Chakraborty (1996) under suitable conditions that the asymptotic generalized variance of the transformation retransformation median regression (i.e. TREMMER) estimate of  $\mathbf{\Gamma}$  in the linear model (1) is minimized by choosing  $\alpha$  to minimize the determinant of the matrix

$$\mathbf{V}(\alpha) = \{\mathbf{J}(\alpha)\}^{-1}\{\mathbf{D}(\alpha)\}\{\mathbf{J}(\alpha)\}^T\}^{-1} . \quad (3)$$

Note that such a selection of  $\alpha$  does not require any knowledge of the form of the density  $f$ , and there is a nice and intuitively appealing geometric interpretation for such an approach. The determinant of  $\mathbf{V}(\alpha)$  is minimized when the columns of  $\Sigma^{-1/2}\mathbf{E}(\alpha)$  are orthogonal to one another. Hence, an alternative way of selecting  $\mathbf{E}(\alpha)$  to achieve a similar goal will be to minimize the ratio of the trace and the  $d$ -th root of the determinant of the matrix  $\{\mathbf{E}(\alpha)\}^T\Sigma^{-1}\mathbf{E}(\alpha)$ , which is equivalent to minimizing the ratio of the arithmetic mean and the geometric mean of the eigenvalues of the positive definite matrix. In the absence of any other better and practically

feasible procedure, we intend to use this criterion for choosing the transformation matrix for our multivariate rank regression. In other words, our recommendation amounts to transforming the response vectors using a new data driven co-ordinate system determined by the transformation matrix  $\mathbf{E}(\alpha)$  such that the co-ordinate system is as orthogonal as possible in the  $d$ -dimensional vector space, where the inner product and orthogonality are defined based on the positive definite dispersion matrix  $\Sigma$  of the residual distribution associated with the linear model (1) [see also Chakraborty and Chaudhuri (1996, 1997) and Chakraborty, Chaudhuri and Oja (1997)]. Of course we need an appropriate estimate of  $\Sigma$  in order to implement such a strategy, and we can get that from the residuals computed at an initial stage after fitting the linear model to the data by any simple and suitable method. Note that it is important that such an estimate of  $\Sigma$  be equivariant under linear transformation of the response vectors.

## 2.2 Multivariate rank regression using Wilcoxon's score

Let us now consider the dispersion functions associated with well known Wilcoxon's rank scores. Such dispersion functions can be expressed in the form

$$\mathcal{D}(\Lambda) = \sum_{1 \leq r < s \leq n} \sum_{r, s \notin \alpha} \left| (\mathbf{Z}_r^{(\alpha)} + \mathbf{Z}_s^{(\alpha)}) - \Lambda(\mathbf{X}_r + \mathbf{X}_s) \right| \quad (4)$$

or

$$\mathcal{D}(\Lambda) = \sum_{1 \leq r < s \leq n} \sum_{r, s \notin \alpha} \left| (\mathbf{Z}_r^{(\alpha)} - \mathbf{Z}_s^{(\alpha)}) - \Lambda(\mathbf{X}_r - \mathbf{X}_s) \right|, \quad (5)$$

where for a  $d$ -dimensional vector  $\mathbf{x} = (x_1, x_2, \dots, x_d)$ ,  $|\mathbf{x}|$  = the  $l_1$ -norm of  $\mathbf{x} = |x_1| + |x_2| + \dots + |x_d|$ . Note that the dispersion in (4) originates from Wilcoxon's signed rank score used in single sample location problems while that in (5) is related to the two sample Wilcoxon's rank test. The second dispersion can also be viewed as a form of Gini's mean difference of multivariate residuals, and it is meaningful to use this dispersion function when there is no intercept term present in the linear model (1). On the other hand the dispersion function in (4) is useful in multivariate linear models with intercept terms. Readers are referred to Aubuchon and Hettmansperger (1989) and Chaudhuri (1992b) for a detailed discussion of these dispersion functions and their use in rank regression in linear models with univariate response.

The estimates of the coefficient matrix obtained through minimization of dispersion functions in (4) and (5) can be viewed as natural extensions of the well-known Hodges-Lehmann estimates from one and two sample location problems into multivariate linear models. Observe at this point

that minimization of any of these two dispersions leads to a co-ordinatewise least absolute deviations problem, and hence the computation of the transformation retransformation estimate  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  can be easily handled by some straight forward modification of the TREMMER algorithm developed by Chakraborty (1996). One only needs to replace the original data by their pairwise averages or differences (depending on whether (4) or (5) is used) before invoking TREMMER. We now state a result that establishes asymptotic optimality of our procedure for choosing the transformation matrix  $\mathbf{E}(\alpha)$  as described in Section 2.1 when rank regression is performed using Wilcoxon's score in a multivariate linear model with the residual having multivariate normal distribution.

**Result 3** *Suppose that the residuals  $\mathbf{e}_i = \mathbf{Y}_i - \mathbf{\Gamma}\mathbf{X}_i$  for  $1 \leq i \leq n$  are i.i.d and have a common  $d$ -variate normal distribution with zero mean and  $\Sigma$  as their common dispersion matrix that does not depend on the regressor (i.e. we have perfect homoscedasticity), and the i.i.d random regressors  $\mathbf{X}_i$ 's have a distribution with an associated  $p \times p$  expected information matrix  $E(\mathbf{X}_i\mathbf{X}_i^T) = \mathbf{Q}$  that is positive definite ensuring asymptotic normality of the co-ordinatewise rank regression estimates obtained using the dispersion function (4) or (5) [cf. the asymptotic results in Chaudhuri (1992b)]. Then our procedure for choosing the set of indices  $\alpha$  and the associated transformation matrix  $\mathbf{E}(\alpha)$  described in Section 2.1 yields a transformation retransformation estimate  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  such that the asymptotic generalized variance of  $n^{1/2}(\hat{\mathbf{\Gamma}}_n^{(\alpha)} - \mathbf{\Gamma})$  tends to its minimum possible value as  $n$  tends to infinity.*

**Proof:** Once again let us fix  $\alpha$  and argue conditionally give the  $(\mathbf{X}_i, \mathbf{Y}_i)$ 's with  $i \in \alpha$ . Note that when the dispersion function (5) is used, there are no intercept terms in the multivariate linear model, and without loss of generality we can assume in this case that the  $\mathbf{X}_i$ 's have zero mean. Under the conditions assumed in the statement of the result, it is easy to establish a Bahadur type asymptotic linear representation of  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  using the asymptotic results in Chaudhuri (1992b), and this implies that as  $n$  tends to infinity, the limiting distribution of  $n^{1/2}(\hat{\mathbf{\Gamma}}_n^{(\alpha)} - \mathbf{\Gamma})$  is multivariate normal with zero mean and a variance covariance matrix that has the form

$$\Sigma^{1/2}\{\mathbf{J}(\alpha)\}^{-1}\mathbf{H}(\alpha)\{[\mathbf{J}(\alpha)]^T\}^{-1}\Sigma^{1/2} \otimes \mathbf{Q}^{-1}, \quad (6)$$

where  $\otimes$  denotes the usual Kornecker product of matrices. Here  $\mathbf{J}(\alpha)$  is the matrix whose rows are obtained by normalizing the rows of the matrix  $\{\Sigma^{-1/2}\mathbf{E}(\alpha)\}^{-1}$  as described in Section 2.1, and  $\mathbf{H}(\alpha)$  is the  $d \times d$  symmetric matrix with  $(i, j)$ -th element equal to  $2\sin^{-1}(\gamma_{ij}/2)$ ,  $\gamma_{ij}$  being the

Euclidean inner product between the  $i$ -th and the  $j$ -th row of  $\mathbf{J}(\alpha)$ . It is the multivariate normality of the residual distribution in the linear model that enables us to simplify the form of the asymptotic dispersion matrix in this special case. It is clear from (6) that the asymptotic generalized variance of the transformation retransformation rank regression estimate will be minimized if we choose  $\alpha$  to minimize  $\det\{\mathbf{H}(\alpha)\}/[\det\{\mathbf{J}(\alpha)\}]^2$ , and this is accomplished when the rows of  $\mathbf{J}(\alpha)$  or equivalently the columns of  $\Sigma^{-1/2}\mathbf{E}(\alpha)$  are orthogonal to one another.  $\square$

### 3 Numerical results: simulation and data analysis

In an attempt to investigate the performance of transformation retransformation rank reregression methodology in finite sample situations, we ran a simulation study and analyzed a couple of real data sets for which there are some appropriate multi-response linear models. We compared our approach with more traditional procedures some of which are not affine equivariant, and as we will gradually see the results turned out to be quite encouraging and favorable for our affine equivariant rank regression.

**A Simulation Study:** We considered a problem with sample size  $n = 30$ , where the data was generated from a multivariate linear model like (1) with  $d = p = 2$ , and the first co-ordinate of  $\mathbf{X}$  was taken to be the constant 1.0 while the second co-ordinate was generated from a standard normal distribution. We chose  $\mathbf{\Gamma}$  as the  $2 \times 2$  zero matrix, and for the random residual, we used three different elliptically symmetric distributions i.e. distributions having densities of the form  $\{\det(\Sigma)\}^{-1/2}f(\mathbf{e}^T\Sigma^{-1}\mathbf{e})$ . These distributions are bivariate normal, bivariate Laplace [i.e. when  $f(\mathbf{e}^T\mathbf{e}) = (2\pi)^{-1}\exp(\sqrt{\mathbf{e}^T\mathbf{e}})$ ] and bivariate  $t$  with 3 degrees of freedom. We used the dispersion function (4) for computing the transformation retransformation estimate  $\hat{\mathbf{\Gamma}}_n^{(\alpha)}$  after choosing  $\alpha$  using the selection procedure described in Section 2.1. Let  $\mathcal{E}_{ols}$  and  $\mathcal{E}_{lad}$  denote the efficiencies of our estimates compared with the ordinary least squares and co-ordinatewise least absolute deviations estimates respectively. These efficiencies were computed using the fourth root of the ratio of the generalized variances of competing estimates [see e.g. Bickel (1964)], and the generalized variances were estimated using 3000 Monte Carlo replications in each case. Since both of ordinary least squares estimate and our estimate of  $\mathbf{\Gamma}$  are affine equivariant,  $\mathcal{E}_{ols}$  does not depend on  $\Sigma$ . We observed that for bivariate normal  $\mathcal{E}_{ols} = 82\%$ , and for bivariate Laplace  $\mathcal{E}_{ols} = 101\%$ . However, for the  $t$  distribution with 3 degrees of freedom, which is a distribution with a fairly heavy tail, we observed that  $\mathcal{E}_{ols} = 150\%$ . Since the co-ordinatewise

least absolute deviations regression does not lead to an affine equivariant estimate of  $\Gamma$ ,  $\mathcal{E}_{lad}$  depends on  $\Sigma$ . For our simulation study, we have used different choices of  $\Sigma$ , and each choice had both diagonal entries equal to 1.0 and both off diagonal entries equal to  $\rho$ . Five different values of  $\rho$  were used, and they are 0.75, 0.80, 0.85, 0.90 and 0.95. The results are summarized in the following table.

Table 3.1: Values of  $\mathcal{E}_{lad}$  for different choices of the residual distribution and  $\rho$ .

| Residual<br>Distribution  | Values of $\rho$ |      |      |      |      |
|---------------------------|------------------|------|------|------|------|
|                           | 0.75             | 0.80 | 0.85 | 0.90 | 0.95 |
| Bivariate Normal          | 158%             | 174% | 185% | 207% | 244% |
| Bivariate Laplace         | 125%             | 128% | 144% | 159% | 193% |
| Bivariate $t$ with 3 d.f. | 121%             | 128% | 141% | 158% | 189% |

**Analysis of Blood Pressure Data:** This data was collected by the Biological Sciences Division of Indian Statistical Institute, Calcutta, and it consists of systolic and diastolic blood pressures of 40 Marwari females residing at Burrabazar area of Calcutta and their ages. It is well known to physiologists that arterial pressure increases with age, and age is considered to be a factor of prime importance in deciding what should be the normal arterial pressure of an individual. As one would expect, there is ample evidence in the data [see e.g. Chakraborty (1996) who analyzed the same data using TREMMER] for the the presence of high positive correlation between systolic and diastolic pressures, and hence one can argue in favor of using an affine equivariant procedure, which is expected to be statistically more efficient for analyzing this data set than a non-equivariant procedure such as the co-ordinatewise least absolute deviations regression. We applied our affine equivariant rank regression procedure based on the dispersion function (4) to this data and obtained the following estimated linear equations : *systolic pressure* =  $100.64 + 0.8(\text{age})$ , and *diastolic pressure* =  $74.04 + 0.32(\text{age})$ . Following Chakraborty (1996), we estimated the sampling variations using 2000 bootstrap samples for each of the competing procedures and observed 66.9% gain in statistical efficiency when our affine equivariant rank regression was compared with co-ordinatewise least absolute deviations regression. The coefficients of age in both the equations here are slightly larger than those obtained by Chakraborty (1996) using TREMMER, and their standard errors (0.20 and 0.11 for systolic and diastolic pressures respectively) estimated through bootstrap turned out to be smaller than those for the TREMMER estimates [cf. Chakraborty (1996)]. It will be appropriate to note here that for using TREMMER, Chakraborty

(1996) reported 14.5% gain in statistical efficiency over non-equivariant co-ordinatewise least absolute deviations.

**Analysis of Demographic Data:** This second data set consists of total fertility rates (TFR), infant mortality rates (IMR) and female literacy rates (FLR) for the years 1971, 1981 and 1991 for sixteen most populated states in India. The data is given in a nicely compiled form in Srinivasan (1995). TFR is defined as the number of children born to a woman in her entire reproductive span assuming that she experiences the level of age-specific fertility rates in a given period of time, and IMR is the number of deaths of infants (i.e. children below age of one year) per thousand live births during a given period. Socio-demographic studies have strongly revealed education of women as a major determinant of visible decline in infant mortality and total fertility levels of the population. Our main interest here is in exploring the nature of dependence of TFR and IMR on FLR as well as the changes in TFR and IMR over time and their regional variations, and for this we have used a multivariate analysis of covariance type model with four regional effect parameters (corresponding to northern, southern, eastern and western regions of the country) and two covariates, namely FLR and time. Once again in view of strong correlation between TFR and IMR [see Chakraborty (1996)], any non-equivariant estimation procedure is expected to perform poorly in this case. Since here one is interested in the differences between regional effects, the dispersion function in (5) is quite appropriate. When we compared our affine equivariant procedure with non-equivariant co-ordinatewise rank regression based on Wilcoxon's score using bootstrap estimates of sampling variations, we observed about 8% gain in statistical efficiency. As in the preceding example, here too we used 2000 bootstrap samples for each competing procedure. In the case of our affine equivariant procedure, time with estimated coefficients -0.4929 and -9.5899 having standard errors 0.1964 and 4.5880 respectively appeared to be a statistically significant covariate indicating decline in both of TFR and IMR over time. FLR too turned out to be a statistically significant covariate with estimated coefficients -0.03775 and -1.3006 having standard errors 0.01223 and 0.2983 respectively indicating a strong influence of female education on decreasing TFR and IMR. However, our analysis did not reveal any statistically significant regional difference in fertility and mortality rates. These findings are in conformity with the results reported in Chakraborty (1996) who analyzed the same data using TREMMER.

## 4 Concluding remarks and discussion

An important issue that emerges at this point is that the problem of multi-

variate rank regression is intrinsically related to the problem of multivariate quantiles and ranks. Readers are referred to Chaudhuri (1996) and Mottonen and Oja (1995) for a detailed review of various notions of multivariate quantiles and ranks. A particularly interesting alternative to our present approach of multivariate rank regression will be to use the ranks associated with spatial or geometric quantiles [see Chaudhuri (1996), Mottonen and Oja (1995)]. Affine equivariance can still be achieved through data driven transformation and retransformation as has been done in Chakraborty, Chaudhuri and Oja (1997), where an equivariant modification of spatial median and an invariant modification of angle test were proposed and studied based on the idea of transformation and retransformation. However, such a geometric concept of quantiles leads to vector valued ranks that are very different in nature from co-ordinatewise ranks, and one needs to redefine the multivariate analog of Jaeckel's dispersion function appropriately using some suitable notion of score functions defined for vector valued ranks.

So far we are able to prove asymptotic optimality of our procedure for selecting the subset of indices  $\alpha$  and the associated transformation matrix  $\mathbf{E}(\alpha)$  only in a very special case, i.e. for dispersions associated with Wilcoxon's rank scores, and when the residual in the linear model (1) is normally distributed. In the case of multivariate median (or least absolute deviations) regression, Chakraborty (1996) was able to show that the proposed data based selection rule for choosing the transformation matrix leads to an asymptotically optimal solution whenever the residual distribution is elliptically symmetric. The nice geometric interpretation of this selection procedure described in Section 2.1 makes us believe that its asymptotic optimality holds under much weaker and more general conditions than what we have assumed in Result 3.

Rank regression in linear models with univariate response generated fascinating research problems and innovative statistical tools for nearly three decades [see Draper (1988)]. This enriched our theoretical understanding of linear model analysis and enabled us to invent new methodology for exploring relationships present among different variables in the data. Multivariate rank regression is likely to lead us to a more fertile ground for methodological and theoretical research. As multi-response problems do arise often in practice, there seems to be a real need for a serious and extensive research of rank regression to deal with such problems.

## Acknowledgements

Research presented here was partially supported by a grant from Indian



Statistical Institute.

## References

- [1] Adichie, J. N. (1967). Estimates of regression parameters based on rank tests. *Ann. Math. Statist.* **38**, 894–904.
- [2] Adichie, J. N. (1978). Rank tests of subhypotheses in the general linear regression. *Ann. Statist.* **6**, 1012–1016.
- [3] Aubuchon, J. C. and Hettmansperger, T. P. (1989). Rank based inference for linear models : asymmetric errors. *Statistics & Probability Letters* **8** 97–107.
- [4] Bai, Z. D., Chen, N. R., Miao, B. Q. and Rao, C. R. (1990). Asymptotic theory of least distances estimate in multivariate linear models. *Statistics* **21**, 503–519.
- [5] Bickel, P. J. (1964). On some alternative estimates for shift in the  $p$ -variate one sample problem. *Ann. Math. Statist.* **35**, 1079–1090.
- [6] Brown, B. M. (1983). Statistical use of spatial median. *J. R. Statist. Soc. B* **45**, 25–30.
- [7] Brown, B. M. and Hettmansperger, T. P. (1987). Affine invariant rank methods in the bivariate location model. *J. R. Statist. Soc. B* **49**, 301–310.
- [8] Brown, B. M. and Hettmansperger, T. P. (1989). An affine invariant bivariate version of the sign test. *J. R. Statist. Soc. B* **51**, 117–125.
- [9] Chakraborty, B. (1996). On multivariate median regression. Technical Report No. 9/96, Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, Calcutta.
- [10] Chakraborty, B. and Chaudhuri, P. (1996). On a transformation and retransformation technique for constructing affine equivariant multivariate median. In *Proc. of the American Mathematical Society*, Vol. 124, pp. 2539–2547.
- [11] Chakraborty, B. and Chaudhuri, P. (1997). On an adaptive transformation and retransformation estimate of multivariate location. *J. R. Statist. Soc. B*. To appear.
- [12] Chakraborty, B., Chaudhuri, P. and Oja, H. (1997). Operating transformation retransformation on spatial median and angle test. *Statistica Sinica*. To appear.
- [13] Chaudhuri, P. (1992a). Multivariate location estimation using extension of  $R$ -estimates through  $U$ -statistics type approach. *Ann. Statist.* **20**, 897–916.
- [14] Chaudhuri, P. (1992b). Generalized regression quantiles : forming a

- useful toolkit for robust linear regression. In *L<sub>1</sub> Statistical Analysis and Related Methods*, Ed. Y. Dodge, pp. 169–185. Amsterdam: North Holland.
- [15] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Am. Statist. Assoc.* **91**, 862–872.
  - [16] Draper, D. (1988). Rank based robust analysis of linear models I : exposition and review (with discussion). *Statistical Science* **3**, 239–271.
  - [17] Davis, J. B. and McKean, J. W. (1993). Rank based methods for multivariate linear models. *J. Am. Statist. Assoc.* **88**, 245–251.
  - [18] Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika* **35**, 414–415.
  - [19] Hettmansperger, T. P. and McKean, J. W. (1977). A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics* **19**, 275–284.
  - [20] Hettmansperger, T. P. and McKean, J. W. (1978). Statistical inference based on ranks. *Psychometrika* **43**, 69–79.
  - [21] Hettmansperger, T. P. and McKean, J. W. (1983). A geometric interpretation of inferences based on ranks in the linear model. *J. Am. Statist. Assoc.* **78**, 885–893.
  - [22] Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. *Ann. Math. Statist.* **43**, 1449–1458.
  - [23] Jureckova, J. (1971). Nonparametric estimation of regression coefficients. *Ann. Math. Statist.* **42**, 1328–1338.
  - [24] Jureckova, J. (1973). Central limit theorem for Wilcoxon rank statistics process. *Ann. Statist.* **1**, 1046–1060.
  - [25] Koenker, R. and Portnoy, S. (1990). *M*-estimation of multivariate regression. *J. Am. Statist. Assoc.* **85**, 1060–1068.
  - [26] Koul, H. L. (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. *Ann. Math. Statist.* **40**, 1950–1979.
  - [27] Koul, H. L. (1970). A class of ADF tests for subhypotheses in multiple linear regression. *Ann. Math. Statist.* **41**, 1273–1281.
  - [28] Lehmann, E. L. (1963a). Robust estimation in analysis of variance. *Ann. Math. Statist.* **34**, 957–966.
  - [29] Lehmann, E. L. (1963b). Asymptotically nonparametric inference : an alternative approach to linear models. *Ann. Math. Statist.* **34**, 1494–1506.
  - [30] Lehmann, E. L. (1964). Asymptotically nonparametric inference in some linear models with one observation per cell. *Ann. Math. Statist.* **35**, 726–734.
  - [31] Mottonen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparam. Statist.* **5**, 201–213.

- [32] Puri, M. L. and Sen, P. K. (1969). A class of rank order tests for a general linear hypothesis. *Ann. Math. Statist.* **40**, 1325–1343.
- [33] Puri, M. L. and Sen, P. K. (1973). A note on asymptotically distribution free tests for subhypotheses in multiple linear regression. *Ann. Statist.* **1**, 553–556.
- [34] Rao, C. R. (1988). Methodology based on the  $L_1$ -norm in statistical inference. *Sankhyā A* **50**, 289–313.
- [35] Srinivasan, K. (1995). Recent fertility trends and prospects in India. *Current Science* **69**, 577–586.

# Multivariate density estimation by probing depth

**Ricardo Fraiman**

*Universidad de la Republica, Montevideo, Uruguay*

**Regina Y. Liu**

*Rutgers University, Piscataway, USA*

**Jean Meloche**

*The University of British Columbia, Vancouver, Canada*

*Abstract:* In high dimension, the estimation of a density is difficult because the observed data gets increasingly sparse with the dimension. This is known as the curse of dimensionality. For that reason, in high dimension, universally consistent estimators such as the kernel density estimator are not practical. In this paper, we consider a class of multivariate densities, within which a density function  $f$  can be expressed as  $f = g \circ D$  for some given notion of data depth  $D$  and some real function  $g$ . We propose a density estimator which is shown to be consistent within the class, and it converges at the same rate as the *univariate* kernel density estimator.

*Key words:* Multivariate density estimation, Data depth, rate of convergence.

AMS subject classification: 62G07, 62H12.

## 1 Introduction

Let  $X_1, \dots, X_n$  be an i.i.d. sample from an unknown density  $f : \mathcal{R}^p \rightarrow [0, \infty)$ . When  $p$  is large, a kernel density estimator of the form

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h^p} K\left(\frac{x - X_i}{h}\right)$$

where  $h$  is the smoothing parameter, is impractical because of the extremely large number of observations needed to “fill” the  $p$ -dimensional space in order to ensure sufficient observations in each “bin” of the kernel. This phenomenon of sparsity of data in high dimensional space is referred to as the “curse of dimensionality” in Bellman (1961).

Many approaches have been developed in the literature in order to address this problem, and in particular projection pursuit techniques (Friedman, Stuetzle and Schroeder, 1984). Projection pursuit avoids this problem by working in low-dimensional linear projections. However, as pointed by Huber (1985), projection pursuit is poorly suited to deal with highly non-linear structures. For these reasons, the analysis of high-dimensional data sets is often made under some additional restrictions. One common practice is to assume that the density  $f$  belongs to some parametric family, so the estimation of the density amounts to the estimation of finitely many parameters. For example, if the underlying density is normal, then one only needs to estimate the mean and the variance. In the same spirit but without the firm grip of parametric assumptions, the so-called “tailor-design density estimates” (cf. Devroye, 1987) are designed to perform well for a particular class of densities. This class can but does not have to be parametric. In general “tailor-design density estimates” are not *universally* consistent, since they are tailored to suit a specific target class of densities. A typical example is the Grenander estimator (Grenander, 1956) which concerns only monotone densities.

In this paper we rely on the general nonparametric smoothing principle to provide a multivariate density estimator, with the idea of enlarging the *neighbourhood* for smoothing so as to include sufficiently many data points even when the dimension is high. Roughly speaking, our approach may be viewed as a generalized version of the following simple nonparametric density estimator

$$\hat{f}_h(x; A_h) = \frac{\frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \in A_h(x))}{m_p(A_h(x))}$$

where  $\mathbf{1}$  denotes the indicator function,  $m_p$  denotes the  $p$ -dimensional Lebesgue measure and  $A_h(x) = \{t \ni \|x - t\| < h\}$ . The above estimate takes advantage of the smoothness of the unknown density  $f$ , and assumes that  $f$  is nearly constant in the neighbourhood  $A_h(x)$ . As indicated above, the difficulty with this approach in  $p$  dimensions is that the volume of the neighbourhood  $A_h(x)$  decreases rapidly with  $p$ . As a result, the variance of the estimator increases rapidly with  $p$  and one is forced to increase the bandwidth  $h$  to obtain a balance between the variance and the squared bias of the estimator. On the other hand,  $A_h(x)$  is not the only set over which

$f$  can be presumed constant. The density  $f$  is also nearly constant over the larger set  $B_h(x) = \{t \ni \|f(x) - f(t)\| < h\}$ . Used as a neighbourhood,  $B_h(x)$  has a large volume even in high dimension, so one is not forced to use a large bandwidth to obtain a balance between the variance and the squared bias. In contrast with the neighbourhood  $A_h(x)$  which does not depend on  $f$ , the neighbourhood  $B_h(x)$  does and needs to be estimated. Of course, the estimation of the neighbourhood  $B_h(x)$  is difficult and may very well offset the improvement resulting from enlarged neighbourhoods.

Our estimator is based on a *neighbourhood* which is between the above two extreme cases corresponding to the sets  $A_h(x)$  and  $B_h(x)$ . Assume that  $f = g \circ D$  for some  $g : \mathcal{R} \rightarrow [0, \infty)$  and some transformation  $D : \mathcal{R}^p \rightarrow [0, \infty)$  that may depend on  $f$ . Under this restriction,  $f$  is constant whenever  $D$  is constant so that  $f$  is nearly constant over the set  $C_h(x) = \{t \ni \|D(x) - D(t)\| < h\}$ . The estimator we propose essentially amounts to using the set  $C_h(x)$  as the neighbourhood.

In recent years, the class of ellipsoidal densities  $f = g \circ D$  with  $D(x) = (x - \mu)^T \Sigma^{-1} (x - \mu)$  for some function  $g$  has received considerable attention because it enables an analysis of multivariate data which does not rely on the validity of the classical multivariate normal theory. The estimator we propose goes hand in hand with such developments, providing an estimate of the density that outperforms other density estimators within that class. The class of ellipsoidal densities is one example among other possible generalisations of the classical multivariate normal family. Various possibilities will be discussed in what follows.

Let  $m_d$  denote the  $d$ -dimensional Lebesgue measure. Throughout the paper, we will assume that the measure  $m_p \circ D^{-1}$  is absolutely continuous with respect to  $m_1$  and  $\mathcal{L}_D$  will denote the Radon Nykodym derivative of  $m_p \circ D^{-1}$  with respect to  $m_1$ . Under this assumption,

$$\Pr\{D(X_1) \in A\} = \int \mathbf{1}\{D(t) \in A\} g(D(t)) dt = \int_A g(r) \mathcal{L}_D(r) dr$$

showing that  $D(X_1)$  has a density, which we will denote  $f_D$ , and that  $f_D = g\mathcal{L}_D$ . The relationship  $f_D = g\mathcal{L}_D$  can also be expressed as

$$f(x) = g(D(x)) = \frac{f_D(D(x))}{\mathcal{L}_D(D(x))}.$$

Thus, if  $D$  is a known transformation, we can estimate  $f(x)$  by

$$\tilde{f}(x; D, h) = \frac{\hat{f}_D(D(x))}{\mathcal{L}_D(D(x))} = \frac{1}{n} \sum_{i=1}^n \frac{K_h(D(x) - D(X_i))}{\mathcal{L}_D(D(x))}.$$

**Example 1** Consider the case of a spherical density  $f$ . This assumption is equivalent to the assumption  $f = g \circ D$  for  $D(x) = \|x\|^2$  and some  $g : \mathcal{R} \rightarrow \mathcal{R}$ . Here,  $D$  is a fixed transformation and does not need to be estimated. It is easy to see that

$$\mathcal{L}_D(r) = \frac{\pi^{p/2}}{\Gamma(p/2)} r^{p/2-1}$$

is the Radon Nykodym derivative of  $m_p \circ D^{-1}$  with respect to  $m_1$ . Thus, in the case of a spherical density  $f$ , we propose the estimator

$$\tilde{f}(x; D, h) = \frac{\Gamma(p/2)}{\pi^{p/2}} (D(x))^{1-p/2} \frac{1}{n} \sum_{i=1}^n K_h(D(x) - D(X_i)).$$

The estimator  $\tilde{f}(x; D, h)$  basically amounts to an estimator of the univariate density  $f_D$  so that we would expect a one-dimensional nonparametric rate of convergence to  $f(x)$  for  $\tilde{f}(x; D, h)$ .

In general,  $D$  could depend on  $f$  and would then need to be estimated. If  $\hat{D}$  is an estimator of  $D$ , we can estimate  $f(x)$  by

$$\tilde{f}(x; \hat{D}, h) = \frac{\hat{f}_{\hat{D}}(\hat{D}(x))}{\mathcal{L}_{\hat{D}}(\hat{D}(x))} = \frac{1}{n} \sum_{i=1}^n \frac{K_h(\hat{D}(x) - \hat{D}(X_i))}{\mathcal{L}_{\hat{D}}(\hat{D}(x))}.$$

**Example 2** Consider the case of an ellipsoidal density  $f$ . This assumption is equivalent to the assumption  $f = g \circ D$  for  $D(x) = (x - \mu)^T \Sigma^{-1} (x - \mu)$  and some  $g : \mathcal{R}^p \rightarrow \mathcal{R}$ . Here,  $D$  depends on unknown parameters  $\mu$  and  $\Sigma$  that need to be estimated. It is easy to see that

$$\mathcal{L}_D(r) = \sqrt{|\Sigma|} \frac{\pi^{p/2}}{\Gamma(p/2)} r^{p/2-1}.$$

is the Radon Nykodym derivative of  $m_p \circ D^{-1}$  with respect to  $m_1$ . Thus, in the case of an ellipsoidal density  $f$  we propose the estimator

$$\tilde{f}(x; \hat{D}, h) = \frac{\Gamma(p/2)}{\sqrt{|\hat{\Sigma}|} \pi^{p/2}} (\hat{D}(x))^{1-p/2} \frac{1}{n} \sum_{i=1}^n K_h(\hat{D}(x) - \hat{D}(X_i))$$

where  $\hat{D}(x) = (x - \hat{\mu})^T \hat{\Sigma}^{-1} (x - \hat{\mu})$  and  $\hat{\mu}$  and  $\hat{\Sigma}$  are some estimates of  $\mu$  and  $\Sigma$ . The transformation  $D$  involves parameters  $\mu$  and  $\Sigma$  for which there exist estimates converging at speed  $1/\sqrt{n}$ . Since the rate of convergence for  $\hat{D}$  is less than  $1/\sqrt{n}$ , we expect the asymptotic behaviour of  $\tilde{f}(x; \hat{D}, h)$  to be unaffected by the estimation of  $D$  and a one-dimensional nonparametric rate of convergence to  $f(x)$  for  $\tilde{f}(x; \hat{D}, h)$ .

Stute and Werner (1991) focus on this last case and propose an estimate of the density based on the above formula, using independent estimates of  $\mu$  and  $\Sigma$  (based on the data of some preliminary sample).

The above two examples are unusual in that an explicit expression for  $\mathcal{L}_D$  is available. When deriving exact expressions for  $\mathcal{L}_D$ , the following relationships are useful. Let  $D$  be some depth. If  $D_{\mu, \Sigma}(x) = D((x - \mu)^T \Sigma^{-1} (x - \mu))$ , we have  $\mathcal{L}_{D_{\mu, \Sigma}}(r) = \sqrt{|\Sigma|} \mathcal{L}_D(r)$ . If  $\rho$  is a monotone transformation and if  $D_\rho(x) = \rho(D(x))$ , we have  $\mathcal{L}_{D_\rho}(r) = (\mathcal{L}_D / |\rho'|)(\rho^{-1}(r))$ . Nevertheless, an explicit expression for  $\mathcal{L}_D$  is usually not available and an approximation for the denominator must be used. Note that provided  $\mathcal{L}_D$  is smooth,

$$\begin{aligned} \bar{\mathcal{L}}_D(D(x)) &= \int K_{h_f}(D(x) - D(t)) dt \\ &= \int K(u) \mathcal{L}_D(D(x) - h_f u) du \rightarrow \mathcal{L}_D(D(x)) \end{aligned}$$

as  $h_f$  converges to 0. Thus, our purpose in this paper is to investigate the properties of

$$\hat{f}(x; D, h) = \frac{1}{n} \sum_{i=1}^n \frac{K_h(D(x) - D(X_i))}{\int K_{h_f}(D(x) - D(t)) dt}$$

and

$$\hat{f}(x; \hat{D}_n, h) = \frac{1}{n} \sum_{i=1}^n \frac{K_h(\hat{D}_n(x) - \hat{D}_n(X_i))}{\int K_{h_f}(\hat{D}_n(x) - \hat{D}_n(t)) dt}$$

where  $K_h(x) = K(x/h)/h$  for some kernel  $K : \mathcal{R} \rightarrow [0, \infty)$  and some bandwidths  $h$  and  $h_f$ . The particular case with  $h_f = 0$  corresponds to the situation where there exist an explicit expression for  $\mathcal{L}_D$  and, since  $h_f$  is merely used to provide a simple approximation for  $\mathcal{L}_D(D(x))$ , our intention is to let  $h_f$  converge to 0 faster than  $h$  does. The estimators  $\hat{f}(x; D, h)$  and  $\hat{f}(x; \hat{D}_n, h)$  are always non-negative but do not integrate to 1 and, in practice, they need to be normalized.

Notions of multivariate depth are interesting candidates for  $D$  because they can usually be estimated at the usual parametric rate  $1/\sqrt{n}$ . Since  $f_D$  is estimated at a one-dimensional nonparametric rate of convergence, the estimation of  $D$  should not affect the overall rate of convergence. This implies that for the class of densities such that  $f = g \circ D$ , we get a one-dimensional nonparametric rate of convergence in a  $p$ -dimensional density estimation problem. This assertion will be proved in the next Section. Many notions of multivariate depth have been defined and studied in the literature (see Small, 1990), including



- Mahalanobis depth (Mahalanobis, 1936)

$$D(x) = 1 / \left( 1 + (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

- Tukey's depth (Tukey, 1975)

$$D(x) = \inf \left\{ \int_H dP(x) \mid H \text{ is a closed half-space containing } x \right\}.$$

- Simplicial depth (Liu, 1990)  $D(x) = \Pr \{x \in S[X_1, \dots, X_{p+1}]\}$  where  $S[X_1, \dots, X_{p+1}]$  is the simplex with vertices  $X_1, \dots, X_{p+1}$ .
- APL depth (Fraiman and Meloche, 1996)  $D(x) = K_\gamma * f(x)$  for some kernel  $K$  and some fixed smoothing parameter  $\gamma$ .

All of the above depths can be estimated at the  $1/\sqrt{n}$  parametric rate so that they can all be estimated at no cost in terms of the asymptotic behaviour of  $\hat{f}(x; \hat{D}_n, h)$ . The depth, however, does have an impact on the asymptotic bias and variance of  $\hat{f}(x; \hat{D}_n, h)$ . More importantly, the depth  $D$  determines the class of densities  $f$  of the form  $f = g \circ D$ . As described in Example 2, for  $D(x) = (x - \mu)^T \Sigma^{-1} (x - \mu)$ ,  $f = g \circ D$  if and only if  $f$  is ellipsoidal. In the case of Simplicial depth, although we know the level curves of  $D$  must be convex, it is not clear how large the class of densities  $f$  of the form  $f = g \circ D$  is. The level curves for APL depth don't even need to be convex but the equation  $f = g \circ D$  does not appear to be satisfied except for ellipsoidal densities.

As noted before, density estimators that perform particularly well under some class of density are called “tailor design density estimate” by Devroye. One can regard the estimator we propose as one that will take advantage of the relationship  $f = g \circ D$  for a given notion of depth. The proposed estimates are not universally consistent (they converge only if  $f = g \circ D$ ) but they provide better performance than the universally consistent estimate on the class of densities  $f$  of the form  $f = g \circ D$ .

## 2 Main Results

In this section, we present results concerning the strong convergence and the asymptotic normality of

$$\hat{f}(x; D, h) = \frac{1}{n} \sum_{i=1}^n \frac{K_h(D(x) - D(X_i))}{\int K_{h_f}(D(x) - D(t)) dt}$$

and

$$\hat{f}(x; \hat{D}_n, h) = \frac{1}{n} \sum_{i=1}^n \frac{K_h(\hat{D}_n(x) - \hat{D}_n(X_i))}{\int K_{h_f}(\hat{D}_n(x) - \hat{D}_n(t)) dt}$$

where  $\hat{D}_n$  is an estimator of  $D$ . We summarize below the assumptions that are needed. Proofs can be found in the Appendix.

**H0:** The bandwidth sequence  $h = h_n$  is such that  $nh^5 \rightarrow \beta^2 \in (0, \infty)$ . The bandwidth sequence  $h_f = h_{nf}$  converges to zero faster than  $h$  does:  $h_f/h \rightarrow 0$ .

**H1:** The kernel  $K$  is symmetric, has a bounded support, integrates to 1 and has three bounded and continuous derivatives.

**H2:**  $X_1, \dots, X_n$  are i.i.d. with some density  $f : \mathcal{R}^p \rightarrow [0, \infty)$  such that  $f = g \circ D$  for some function  $g : \mathcal{R} \rightarrow [0, \infty)$ . Both  $g$  and  $f$  are bounded and have two bounded and continuous derivatives.

**H3:** There exist a 1 – 1 and continuously differentiable transformation  $T : \mathcal{R} \times [0, 1]^{p-1} \rightarrow \mathcal{R}^p$  such that  $D(T(r, \theta)) = r$  for all  $r \in \mathcal{R}$  and all  $\theta \in [0, 1]^{p-1}$ . The transformations  $T$  and its Jacobian  $J_T$  have two bounded and continuous partial derivatives with respect to  $r$ .

**H4:** The inverse image by  $D$  of a bounded set is bounded and  $x$  is in the interior of the support of  $D$ .

Assumptions **H0** and **H1** are more or less the standard assumption for the bandwidth and the kernel in kernel density estimation. The bounded support of the kernel is not usually needed but simplifies the proofs. The necessity of Assumptions **H2** can be explained as follows. Since  $\hat{f}(x; D, h)$  is a function of  $D(x)$  so that we can only hope to get consistency if  $f = g \circ D$  for some  $g : \mathcal{R} \rightarrow [0, \infty)$ . Note that, by virtue of **H3**,

$$\begin{aligned} \Pr\{D(X_1) \in A\} &= \int \mathbf{1}\{D(t) \in A\} g(D(t)) dt \\ &= \int \int \mathbf{1}\{D(T(r, \theta)) \in A\} g(D(T(r, \theta))) |J_T(r, \theta)| d\theta dr \\ &= \int \int \mathbf{1}\{r \in A\} g(r) |J_T(r, \theta)| d\theta dr \\ &= \int_A g(r) \mathcal{L}_D(r) dr \end{aligned}$$

where  $\mathcal{L}_D(r) = \int |J_T(r, \theta)| d\theta$ . Thus, **H3** implies that the random variable  $D(X_1)$  has the density  $f_D = g\mathcal{L}_D$ .

**Theorem 1** *If **H0-H3** hold,*

$$\begin{aligned} \sqrt{nh} \left\{ \hat{f}(x; D, h) - f(x) \right\} &\xrightarrow{\mathcal{L}} \\ N \left( \frac{1}{2} \beta \int u^2 K(u) du \frac{f_D''(D(x))}{\mathcal{L}_D(D(x))}, \int K^2(u) du \frac{f(x)}{\mathcal{L}_D(D(x))} \right). \end{aligned}$$

The asymptotic distribution of  $\hat{f}(x; D, h)$  should be compared to that of the multivariate kernel density estimator  $\hat{f}(x; h)$  which is well known to be

$$\sqrt{nh^d} \left\{ \hat{f}(x; h) - f(x) \right\} \xrightarrow{\mathcal{L}} N \left( \frac{1}{2} \beta \int u^2 K(u) du \nabla^2 f(x), \int K^2(u) du f(x) \right).$$

Table 1 provides the asymptotic bias and variance for both  $\hat{f}(x; h)$  and  $\hat{f}(x; D, h)$ . The most striking difference is the rate of the convergence to 0 for the asymptotic variance. The asymptotic bias has the same rate of the convergence to 0, but the constants depend on  $f$  in different ways. In the table,  $\alpha = \int u^2 K(u) du$  and  $\beta = \int K^2(u) du$ .

Table 1: The asymptotic bias and variance for  $\hat{f}(x; h)$  and  $\hat{f}(x; D, h)$ .

| Estimator                        | Asymptotic Bias                                                  | Asymptotic Variance                                   |
|----------------------------------|------------------------------------------------------------------|-------------------------------------------------------|
| $\hat{f}(x; h)$                  | $\frac{1}{2} \alpha h^2 \nabla^2 f(x)$                           | $\frac{\beta}{nh^d} f(x)$                             |
| $\hat{f}(x; D, h)$               | $\frac{1}{2} \alpha h^2 \frac{f_D''(D(x))}{\mathcal{L}_D(D(x))}$ | $\frac{\beta}{nh} \frac{f(x)}{\mathcal{L}_D^p(D(x))}$ |
| $D(x) =  x - \mu  \quad (p = 1)$ | $\frac{1}{4} \alpha h^2 f_D''(D(x))$                             | $\frac{\beta}{nh} \frac{f(x)}{2}$                     |

Our proof of the asymptotic normality for  $\hat{f}(x; \hat{D}_n, h)$  uses a three term Taylor series approximation for  $\hat{f}(x; \hat{D}_n, h)$ . We prove that provided  $\sqrt{nh} \|\hat{D}_n - D\|_\infty^3 = o_p(h_f^4)$ ,

$$\sqrt{nh} \left\{ \hat{f}(x; \hat{D}_n, h) - \hat{f}(x; D, h) \right\} \xrightarrow{P} 0.$$

Note that for the kernel density estimator, the optimal bandwidth is of the order  $n^{-1/5}$  so that if  $\|\hat{D}_n - D\|_\infty = O_p(1/\sqrt{n})$ , the condition  $\sqrt{nh} \|\hat{D}_n - D\|_\infty^3 = o_p(h_f^4)$  amounts to  $h/(n^2 h_f^8) \rightarrow 0$  and can be satisfied for the optimal bandwidth  $h$  and a slightly faster  $h_f$ . A higher order Taylor series approximation for  $\hat{f}(x; \hat{D}_n, h)$  would result in weaker restrictions on the bandwidths  $h$  and  $h_f$  but at the cost of additional smoothness requirements on  $K$ ,  $T$  and  $J_T$ .

**Theorem 2** Assume **H0-H4** hold and define

$$H_{nk}(s, t) = n \mathbb{E} \left\{ A_n^k(x, X_1) A_n^k(x, X_2) | X_1 = s, X_2 = t \right\}$$

and

$$G_{nk}(s, t) = n \mathbb{E} A_n^k(x, s) A_n^k(x, t)$$

where  $A_n(x, y) = ((\hat{D}_n(x) - \hat{D}_n(y)) - (D(x) - D(y)))$ . If  $H_{nk}$  and  $G_{nk}$  have two continuous and uniformly bounded derivatives for  $k = 1, 2$  and if  $\sqrt{nh} \|\hat{D}_n - D\|_\infty^3 = o_p(h_f^4)$ ,  $\hat{f}(x; \hat{D}_n, h)$  and  $\hat{f}(x; D, h)$  have the same asymptotic distribution.

Besides the smoothness assumption **(H2)** made on  $f$  and  $g$ , the application of Theorem 2 requires the verification of **H3**, the existence of a smooth change of variable transformation  $T$  such that  $D(T(r, \theta)) = r$ . The existence of such a transformation is guaranteed whenever  $D$  has a unique global maximum  $M$  (such a global maximum can be regarded as the deepest point or a median) and is decreasing along every ray originating from  $M$ . In such circumstances, we can define  $T$  as the inverse of the “polar transformation”  $(D, \Theta)$  where  $D$  is replacing the usual norm and where the angles  $\Theta$  are determined about the maximum  $M$ . The smoothness of the resulting  $T$  is equivalent to the smoothness of  $D$ . Thus, we can reformulate all the smoothness assumptions in terms of the smoothness of  $g$  and  $D$ .

### 3 Simulation

In this section, we present a small simulation that compares  $\hat{f}(x; \hat{D}_n, h)$  to the kernel density estimator  $\hat{f}(x, h)$  in cases where the underlying density  $f$  is of the form  $f = g \circ D$ . The simulation involves the Mahalanobis depth. It involves 100 samples of 50 observations i.i.d. with distribution

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{pmatrix}\right).$$

For the Mahalanobis depth,  $D(x, y) = (x - \mu)^T \Sigma^{-1} (x - \mu)$  and  $\hat{D}_n(x, y) = (x - \hat{\mu})^T \hat{\Sigma}^{-1} (x - \hat{\mu})$  with the usual sample moment estimators  $\hat{\mu}$  and  $\hat{\Sigma}$ . In this case we know  $\mathcal{L}_D(r)$  up to a normalizing factor and we use  $h_f = 0$ . We evaluate both  $\hat{f}(x; \hat{D}_n, h)$  and  $\hat{f}(x, h)$  with their respective ASE-optimal bandwidth determined by minimizing over  $h$

$$ASE(\hat{f}(x; \hat{D}_n, h), h) = \frac{1}{50} \sum_{i=1}^{50} \left( \hat{f}(X_i; \hat{D}_n, h) - f(X_i) \right)^2$$

and

$$ASE(\hat{f}(x; h), h) = \frac{1}{50} \sum_{i=1}^{50} \left( \hat{f}(X_i; h) - f(X_i) \right)^2$$

respectively.

Note that the numerator of  $\hat{f}(x; \hat{D}_n, h)$  is in fact a kernel density estimate for the data  $\hat{D}_n(X_1), \dots, \hat{D}_n(X_n)$  evaluated at  $\hat{D}_n(x)$ . In practice, a boundary correction kernel must be used because  $\hat{D}_n$  has a bounded range. It is also possible to use a transformation to avoid having to make a boundary correction. The simulation uses a boundary correction kernel.

Table 2 below summarizes the results. With the Mahalanobis depth,  $\hat{f}(x; \hat{D}_n, h)$  clearly outperforms  $\hat{f}(x, h)$ . The average ASE for  $\hat{f}(x; \hat{D}_n, h)$  is

about three times smaller than the average ASE for  $\hat{f}(x, h)$ . According to the theoretical results of the previous section, for such elliptical densities, the difference between the performance of  $\hat{f}(x; \hat{D}_n, h)$  and  $\hat{f}(x, h)$  should grow with the dimension.

Table 2: ASE of  $\hat{f}(x, h)$  and  $\hat{f}(x; \hat{D}_n, h)$ .

|         | Mahalanobis     |                            |
|---------|-----------------|----------------------------|
|         | $\hat{f}(x, h)$ | $\hat{f}(x; \hat{D}_n, h)$ |
| Average | 0.96            | 0.32                       |
| S.D.    | 0.28            | 0.19                       |

Figure 1 illustrates a very typical outcome of the above simulation. Figure 1 a) displays the 50 observations, Figure 1 b) displays the true density, Figure 1 c) displays the kernel density estimate  $\hat{f}(x, h)$  and Figure 1 d) displays  $\hat{f}(x; \hat{D}_n, h)$  for the Mahalanobis depth. The estimate  $\hat{f}(x; \hat{D}_n, h)$  is by construction perfectly ellipsoidal and the spurious bumps and dips found in  $\hat{f}(x, h)$  have disappeared. Also the bandwidth minimizing the ASE for the kernel density estimate is much smaller than the bandwidth minimizing the ASE for the Mahalanobis depth density estimate.

## 4 Conclusion

In cases where the unknown density  $f$  satisfies  $f = g \circ D$ , our theoretical results show that, in high dimension,  $\hat{f}(x; \hat{D}_n, h)$  has better asymptotic performance than the usual kernel density estimator  $\hat{f}(x, h)$ . Our simulations suggest this is true in two dimension, even for small samples.

The theoretical results also make smoothness assumptions that excludes cases where the unknown density  $f$  is multimodal. Even though the smoothness assumptions can be reduced to include such cases, an important practical problem remains. For multimodal densities,  $f_D$  (estimated by the numerator of  $\hat{f}(x; \hat{D}_n, h)$ ) and  $\mathcal{L}_D$  (estimated by the denominator of  $\hat{f}(x; \hat{D}_n, h)$ ) are discontinuous functions. Depth based estimation for such densities would require the development of a reasonably good density estimator for discontinuous densities.

## Appendix: Proofs of Theorems 1 and 2

This section is devoted to the proof of Theorems 1 and 2.

**Proof of Theorem 1:** Define

$$\hat{f}_D(r; h) = \frac{1}{n} \sum_{i=1}^n K_h(r - D(X_i)),$$

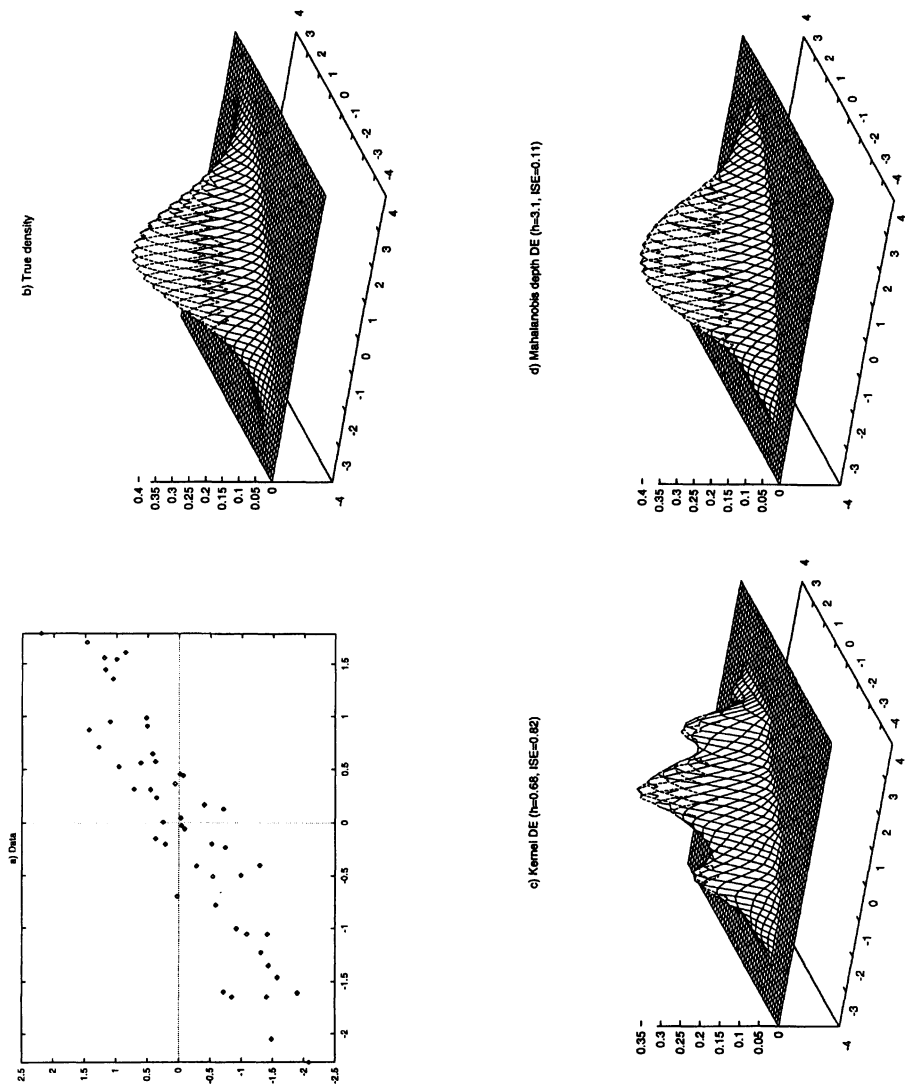


Figure 1: Outcome of the simulation.

$$\bar{\mathcal{L}}_D(r) = \int K_{h_f}(r - D(t)) dt = \int K(u) \mathcal{L}_D(r - h_f u) du$$

and

$$\bar{f}_D(r) = E K_h(r - D(X_1)) = \int K(u) f_D(r - hu) du.$$

With this notation,

$$\hat{f}(x; D, h) = \frac{\hat{f}_D(D(x); h)}{\bar{\mathcal{L}}_D(D(x))}$$

and  $\hat{f}_D(r; h)$  is clearly a kernel density estimator of  $f_D(r)$ . The asymptotic normality of the kernel density estimator is well known so that

$$\sqrt{nh} \left\{ \hat{f}_D(r; h) - \bar{f}_D(r) \right\} \xrightarrow{\mathcal{L}} N \left( 0, f_D(r) \int K^2(u) du \right).$$

Therefore (since  $g = f_D/\mathcal{L}_D$ ),

$$\sqrt{nh} \frac{\hat{f}_D(r; h) - \bar{f}_D(r)}{\bar{\mathcal{L}}_D(r)} \xrightarrow{\mathcal{L}} N \left( 0, \frac{g(r)}{\mathcal{L}_D(r)} \int K^2(u) du \right).$$

The stated asymptotic normality for  $\hat{f}(x; D, h)$  follows because

$$\begin{aligned} \sqrt{nh} \left\{ \frac{\bar{f}_D(r)}{\bar{\mathcal{L}}_D(r)} - \frac{f_D(r)}{\mathcal{L}_D(r)} \right\} &= \sqrt{nh} \frac{\bar{f}_D(r) \mathcal{L}_D(r) - f_D(r) \bar{\mathcal{L}}_D(r)}{\bar{\mathcal{L}}_D(r) \mathcal{L}_D(r)} \\ &= \sqrt{nh} \frac{\int K(u) \{f_D(r-hu) \mathcal{L}_D(r) - f_D(r) \mathcal{L}_D(r-h_f u)\} du}{\bar{\mathcal{L}}_D(r) \mathcal{L}_D(r)} \\ &\rightarrow \frac{1}{2} \beta \frac{f_D''(r)}{\mathcal{L}_D(r)} \int u^2 K(u) du. \end{aligned}$$

**Lemma A** Let  $\gamma_n$  be some sequence that converges to infinity and assume that

$$\gamma_n \left\{ \frac{a_n}{b_n} - c \right\} \xrightarrow{\mathcal{L}} G.$$

If  $\gamma_n(A_n - a_n) \xrightarrow{\mathcal{P}} 0$ ,  $\gamma_n(B_n - b_n) \xrightarrow{\mathcal{P}} 0$  and  $B_n \xrightarrow{\mathcal{P}} b > 0$ , then

$$\gamma_n \left\{ \frac{A_n}{B_n} - c \right\} \xrightarrow{\mathcal{L}} G.$$

**Proof:** Simply write

$$\gamma_n \left\{ \frac{A_n}{B_n} - c \right\} = \gamma_n \left\{ \frac{A_n}{B_n} - \frac{a_n}{b_n} \right\} + \gamma_n \left\{ \frac{a_n}{b_n} - c \right\},$$

note that

$$\gamma_n \left\{ \frac{A_n}{B_n} - \frac{a_n}{b_n} \right\} = \frac{\gamma_n}{B_n} (A_n - a_n) - \frac{\gamma_n}{B_n} (B_n - b_n) \frac{a_n}{b_n}$$

and apply Slutsky Theorem.  $\square$

**Lemma B** If  $\int \left(K^{(2)}(u)\right)^2 du < \infty$  and if the density  $f_D$  of  $D(X_1)$  is bounded,

$$\sup_h h^{2k} \mathbb{E} h \left( K_h^{(k)}(r - D(X_1)) \right)^2 < \infty.$$

**Proof:** Note that

$$\begin{aligned} h^{2k} \mathbb{E} h \left( K_h^{(k)}(r - D(X_1)) \right)^2 &= h^{2k+1} \int \left( K_h^{(k)}(r - t) \right)^2 f_D(t) dt \\ &= h^{2k+1} \int h^{-2(k+1)} \left( K^{(k)}\left(\frac{r-t}{h}\right) \right)^2 f_D(t) dt \\ &= \int \frac{1}{h} \left( K^{(k)}\left(\frac{r-t}{h}\right) \right)^2 f_D(t) dt \\ &= \int \left( K^{(k)}(u) \right)^2 f_D(r - hu) dt \\ &\leq \|f_D\|_\infty \int \left( K^{(k)}(u) \right)^2 dt \quad \square \end{aligned}$$

**Lemma 1** Assume that **H3** holds. If  $G_n : \mathcal{R}^2 \rightarrow \mathcal{R}$  has two continuous and uniformly bounded derivatives, then for  $k \in \{1, 2\}$ ,

$$\sup_n \left| \int \int G_n(x, y) K_h^{(k)}(u - D(x)) K_h^{(k)}(v - D(y)) dx dy \right| < \infty.$$

**Proof:** By virtue of **H3**,

$$\begin{aligned} &\int \int G_n(x, y) K_h^{(k)}(u - D(x)) K_h^{(k)}(v - D(y)) dx dy \\ &= \int \int \int \int G_n(T(s, \theta_1), T(t, \theta_2)) K_h^{(k)}(u - D(T(s, \theta_1))) K_h^{(k)}(v - D(T(t, \theta_2))) \\ &\quad |J_T(s, \theta_1)| |J_T(t, \theta_2)| d\theta_1 d\theta_2 ds dt \\ &= \int \int \bar{G}_n(s, t) K_h^{(k)}(u - s) K_h^{(k)}(v - t) ds dt \end{aligned}$$

with

$$\bar{G}_n(s, t) = \int \int G_n(T(s, \theta_1), T(t, \theta_2)) |J_T(s, \theta_1)| |J_T(t, \theta_2)| d\theta_1 d\theta_2.$$

Since  $T$ ,  $J_T$  and  $G_n$  have two continuous and uniformly bounded derivatives, so does  $\bar{G}_n$  and the result easily follows from

$$\begin{aligned} &\int \int G_n(x, y) K_h^{(k)}(u - D(x)) K_h^{(k)}(v - D(y)) dx dy \\ &= \int \int \bar{G}_n(s, t) K_h^{(k)}(u - s) K_h^{(k)}(v - t) ds dt \\ &= \int \int \frac{\partial^k}{\partial u^k} \frac{\partial^k}{\partial v^k} \bar{G}_n(s, t) K_h(u - s) K_h(v - t) ds dt \\ &= \int \int \frac{\partial^k}{\partial u^k} \frac{\partial^k}{\partial v^k} \bar{G}_n(u - s, v - t) K_h(s) K_h(t) ds dt \\ &\leq \left\| \frac{\partial^k}{\partial u^k} \frac{\partial^k}{\partial v^k} \bar{G}_n(u, v) \right\|_\infty. \quad \square \end{aligned}$$



**Lemma 2** Assume that **H3** holds. If  $H_n : \mathcal{R}^2 \rightarrow \mathcal{R}$  has two continuous and uniformly bounded derivatives, and if  $f$  has two continuous and bounded derivatives, then for  $k \in \{1, 2\}$ ,

$$\sup_n \left| \mathbb{E} H_n(X_1, X_2) K_h^{(k)}(u - D(X_1)) K_h^{(k)}(v - D(X_2)) \right| < \infty.$$

**Proof:** Note that

$$\begin{aligned} & \mathbb{E} H_n(X_1, X_2) K_h^{(k)}(u - D(X_1)) K_h^{(k)}(v - D(X_2)) \\ &= \int \int H_n(x, y) K_h^{(k)}(u - D(x)) K_h^{(k)}(v - D(y)) f(x) f(y) dx dy \end{aligned}$$

and apply Lemma 1 to  $G_n(x, y) = H_n(x, y) f(x) f(y)$ .  $\square$

**Proof of Theorem 2:** First note that  $|A_n(u, v)| \leq 2\|\hat{D}_n - D\|_\infty$  and that  $\sqrt{nh} \|\hat{D}_n - D\|_\infty^3 = o_p(h_f^4)$  implies  $\sqrt{nh} \|\hat{D}_n - D\|_\infty^3 = o_p(h^4)$  and  $\|\hat{D}_n - D\|_\infty = O_p(h)$ . Using Lemma A, we can prove that  $\hat{f}(x; \hat{D}_n, h)$  and  $\hat{f}(x; D, h)$  have the same asymptotic distribution by showing

$$\sqrt{nh} \{ \hat{f}_{\hat{D}}(\hat{D}_n(x); h) - \hat{f}_D(D(x); h) \} \xrightarrow{P} 0 \quad (1)$$

and

$$\sqrt{nh} \{ \bar{\mathcal{L}}_{\hat{D}}(\hat{D}_n(x)) - \bar{\mathcal{L}}_D(D(x)) \} \xrightarrow{P} 0 \quad (2)$$

We use Taylor series approximations for the numerator (1) and the denominator (2) separately. For the numerator,

$$\begin{aligned} \hat{f}_{\hat{D}}(\hat{D}_n(x); h) &= \sum_{k=0}^2 \frac{1}{n} \sum A_n^k(x, X_i) K_h^{(k)}(D(x) - D(X_i)) \\ &\quad + \frac{1}{n} \sum A_n^3(x, X_i) K_h^{(3)}(\hat{\theta}) \end{aligned}$$

for some  $\hat{\theta}$  on the segment joining  $D(x) - D(X_i)$  to  $\hat{D}_n(x) - \hat{D}_n(X_i)$ . For the remainder, we have

$$A_n^3(x, X_i) K_h^{(3)}(\hat{\theta}) \leq 8\|\hat{D}_n - D\|_\infty^3 \frac{\|K^{(3)}\|_\infty}{h^4}$$

and  $\sqrt{nh} \|\hat{D}_n - D\|_\infty^3 = o_p(h^4)$  ensures that the remainder is negligible. Thus, (1) follows from

$$nh \mathbb{E} \left( \frac{1}{n} \sum A_n^k(x, X_i) K_h^{(k)}(D(x) - D(X_i)) \right)^2 \rightarrow 0$$

for  $k = 1, 2$ . The above can be written as a double sum over all indices  $i$  and  $j$ . For  $i = j$ , we have

$$\begin{aligned} & nh \frac{n}{n^2} \mathbb{E} A_n^{2k}(x, X_1) \left( K_h^{(k)}(D(x) - D(X_1)) \right)^2 \\ & \leq h \left( \frac{2\|\hat{D}_n - D\|_\infty}{h} \right)^{2k} \mathbb{E} h^{2k} \left( K_h^{(k)}(D(x) - D(X_1)) \right)^2 \end{aligned}$$

which converges to zero (invoke Lemma B) provided  $\|\hat{D}_n - D\|_\infty = O_p(h)$  (a consequence of  $\sqrt{nh} \|\hat{D}_n - D\|_\infty^3 = o_p(h^4)$ ). When  $i \neq j$ , we have

$$\begin{aligned} & nh \frac{n(n-1)}{n^2} \mathbb{E} A_n^k(x, X_1) K_h^{(k)}(D(x) - D(X_1)) A_n^k(x, X_2) K_h^{(k)}(D(x) - D(X_2)) \\ & = h \mathbb{E} H_{nk}(X_1, X_2) K_h^{(k)}(D(x) - D(X_1)) K_h^{(k)}(D(x) - D(X_2)) \end{aligned}$$

which converges to zero as well because according to Lemma 2,

$$\sup_n |\mathbb{E} H_{nk}(X_1, X_2) K_h^{(k)}(D(x) - D(X_1)) K_h^{(k)}(D(x) - D(X_2))| < \infty.$$

For the denominator, we consider the expansion

$$\bar{\mathcal{L}}_{\hat{D}}(\hat{D}_n(x)) = \sum_{k=0}^2 \int A_n^k(x, t) K_{h_f}^{(k)}(D(x) - D(t)) dt + \int A_n^3(x, t) K_{h_f}^{(3)}(\hat{\theta}_t) dt$$

for some  $\hat{\theta}_t$  on the segment joining  $D(x) - D(t)$  to  $\hat{D}_n(x) - \hat{D}_n(t)$ . Note that since  $K$  has a bounded support, **H4** implies that for  $n$  large enough,  $K_{h_f}^{(3)}(\hat{\theta}_t)$  almost surely vanishes outside of a bounded set so that (2) follows from

$$\sqrt{nh} A_n^3(x, t) K_{h_f}^{(3)}(\hat{\theta}_t) \xrightarrow{\mathcal{P}} 0$$

for almost all  $t$  and

$$nh \mathbb{E} \left( \int A_n^k(x, t) K_{h_f}^{(k)}(D(x) - D(t)) dt \right)^2 \rightarrow 0$$

for  $k = 1, 2$ . The proof of the above is similar to the one we made for the numerator but uses Lemma 1 instead of Lemma 2.  $\square$

## Acknowledgements

This research was supported by NSF Grant DMS90-22126, CONICYT #37 and NSERC 5-81089.

## References

- [1] Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton Univ. Press.

- [2] Brown, B. M. (1983). Statistical uses of the spatial median. *J. R. Statist. Soc. B* **45**, 25-30.
- [3] Devroye, L. (1987). *A Course in Density Estimation*. Progress in Probability and Statistics, Vol. 14, Birkhauser.
- [4] Fraiman, R. and Meloche, J. (1996). Multivariate L-Statistics. Technical Report, The University of British Columbia, Department of Statistics.
- [5] Friedman, J.H. (1991). Multivariate Adaptive Regression Splines. *Ann. Statist.* **19**, 1-67.
- [6] Friedman, J.H., Stuetzle, W. and Schroeder, A. (1984). Projection pursuit density estimation. *J. Am. Statist. Assoc.* **79**, 599-608.
- [7] Grenander, U. (1956). On the theory of mortality measurements, part II. *Skandinavisk Aktuarietidskrift* **39**, 125-153.
- [8] Huber, P.J. (1985). Projection pursuit. *Ann. Statist.* **13**, 435-475.
- [9] Liu, R. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18**, 405-414.
- [10] Mahalanobis, P.C. (1936). On the generalized distance in statistics. In *Proc. of the National Academy of India*, Vol. 12, pp. 49-55.
- [11] Oja, H., (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters* **1**, 327-332.
- [12] Singh, K. (1991). A notion of majority depth. Technical Report, Rutgers University, Department of Statistics.
- [13] Small, C.G. (1990). A survey of multidimensional medians. *Int. Statist. Rev.* **58**, 263-277.
- [14] Stute, W. and Werner, U. (1991). Nonparametric estimation of elliptically contoured densities. In *Nonparametric Functional Estimation and Related Topics*, Ed. G. Roussas, pp. 173-190. NATO ASI Series.
- [15] Tukey, J.W. (1975). Mathematics and picturing data. In *Proc. of the 1974 International Congress of Mathematicians*, Vol. 2, pp. 523-531. Vancouver.

# Modes, caps and concentration: A geometric approach to estimation on the sphere

Philip Milasevic<sup>1</sup>

*University of Lausanne, Switzerland*

Deborah Nolan

*University of California at Berkeley, USA*

*Abstract:* This paper surveys some recent results on mode and concentration estimation in multidimensions, including excess-mass sets and multidimensional quantiles. Extensions of these estimators to the hypersphere are developed here. In particular, the modal direction is measured by the center of a minimal cap, and concentration is measured by a function of the opening of that cap. For samples from a distribution for which the minimal cap is unique, it is shown that the center and the opening of the empirical cap are strongly consistent estimators for their respective parameters. Rates of convergence and limiting distributions of the estimators are established by means of empirical process theory.

*Key words:* Multidimensional mode estimation, directional data, cube-root rates, empirical processes.

AMS subject classification: Primary 62H12; secondary 62H11, 62H10, 60G99.

## 1 Background

A variety of location estimators for multidimensional data have been recently proposed and investigated. Examples of extensions of the median to higher dimensions include the  $L_1$  median (Brown, 1983, and Ducharme and Milasevic, 1987), Oja's simplex (Oja, 1983), the halfspace median (Donoho

---

<sup>1</sup>Milasevic died while this manuscript was in preparation. Nolan would like to dedicate it to his memory.

and Gasko, 1992), and the simplicial depth median (Liu, 1990). For an overview of recent results in this area see Small (1990). Small (1990) also discusses extensions of these notions of the median to directional data, and Liu and Singh (1992) investigate them in greater detail. Asymptotic properties of the simplicial depth were studied by Arcones et. al. (1994) and Dümbgen (1992), those of the halfspace median have been studied by Nolan (1992), and Chaudhuri (1996) has presented a general approach to studying quantiles in multidimensions.

Parallel to the development of the multidimensional median there have been investigations into the properties of estimators of the mode and concentration. Chernoff (1964) and Venter (1967) estimate the mode of a density function in one dimension by the center of the interval of fixed length to contain the greatest number of observations and by the center of the shortest interval to contain at least half of the observations, respectively. Sager (1979) generalized these univariate set statistics to the multidimensional case. He estimates the contours of a unimodal density by a sequence of nested convex sets. The first and largest set is the smallest convex set to contain a fixed proportion  $q$  of the observations; the second set is the smallest convex set that contains proportion  $q$  of the observations within the first set, and so forth. Eddy and Hartigan (1977) proposed a similar multidimensional estimator. The asymptotic properties of the shorth, the shortest interval to contain at least half of the observations, were investigated by Grübel (1988) and Kim and Pollard (1990). Grübel (1988) handled the length of the shorth, and Kim and Pollard developed theory for cube-root rates of convergence to address the center of the shorth. Einmahl and Mason (1992) produced asymptotic theory for generalizations of the length of the shorth.

Close relatives to these estimators of the mode are contour estimators based on excess-mass, proposed independently by Hartigan (1987) and Müller and Sawitzki (1991). An excess mass set for a distribution is the set that maximizes the difference between the probability content of the set and a multiple of its Lebesgue measure.

Nolan (1991) considered the properties of these sets when restricted to ellipsoids, and found the parameters of the ellipsoid have cube-root rates of convergence. Polonik (1995a,b) provides a comprehensive investigation into the properties of these excess mass sets, their connections to maximum likelihood estimation under shape restrictions, and their use in tests of multimodality. He (Polonik, 1995b) shows that the excess mass sets can be used to form a density estimator. The estimator coincides with Grenander's estimator in one dimension when the sets are restricted to intervals with left endpoint 0, and with Sager's estimator in higher dimensions.

## 2 The hypersphere

Here, we consider the extension to directional data of the shorth's intuitive geometric approach to measuring location and concentration. Location and concentration of the distribution are derived from the smallest cap on the sphere that has probability content at least  $\alpha$ , for  $0 < \alpha < 1$ . When unique, the center of this minimal cap can be interpreted as a *modal* direction of the distribution and the cosine of the half-angle of this cap is a measure of *concentration*. Given a sample of size  $n$  on the sphere, the modal direction and concentration can be estimated from the empirical minimal cap, the smallest cap that contains at least  $n\alpha$  observations.

The method proposed here is in some sense a geometric counterpart to Watson's estimator (Watson, 1983, Chapter 5) because it too can provide information on location for axial distributions. For example, with the Scheidegger-Watson, Arnold and uniform distributions, there is no unique minimal cap, but the collection of centers of the caps provides meaningful modal directions. For the Scheidegger-Watson distribution the collection of centers is the axis of symmetry; for the Arnold distribution it is the plane of the girdle and for the uniform distribution it is the entire sphere. The minimal cap differs from Watson's estimator and other current estimators of location and concentration on the hypersphere (Ducharme and Milasevic, 1987, Fisher et. al., 1987, Watson, 1983) in its geometric rather than metric nature. As with other mode estimates, the convergence of the center of the sample minimal cap is at a cube-root rate. The concentration of the cap however has a square-root rate of convergence, as it is similar in behavior to a quantile estimator.

To formally define the minimal cap, we introduce some notation. Let  $S$  denote the Euclidean unit sphere  $S^{p-1}$  in  $R^p$  and  $C(u, t)$  the (hyper)spherical cap with center  $u$  and half-angle  $\arccos(t)$ . That is,  $C(u, t) = \{x \in S : u'x \geq t\}$ , for  $t \in [-1, 1]$ . To simplify notation, given a distribution  $F$  on  $S$  we shall write  $F(u, t)$  for  $F(C(u, t))$ .

**Definition 1** Let  $0 < \alpha < 1$  and  $F$  be a probability measure on  $S$ . A cap  $C(u_0, t_0)$  is called a *minimal  $\alpha$ -cap* of  $F$  if  $F(u_0, t_0) \geq \alpha$  and if for each  $t > t_0$ ,  $\sup_{u \in S} F(u, t) < \alpha$ .

**Definition 2** Given a set  $X^{(n)}$  of  $n$  points on  $S$ ,  $C_n = C(u_n, t_n)$  is called a *minimal  $\alpha$ -cap* of  $X^{(n)}$  if it is a minimal  $\alpha$ -cap of the empirical measure  $F_n$  based on  $X^{(n)}$ .

Note that  $t_0 = \sup\{t : \sup_{u \in S} F(u, t) \geq \alpha\}$ , for  $0 < \alpha < 1$ . We shall call this value the  $\alpha$ -concentration coefficient of  $F$ . If  $C(u_0, t_0)$  is the unique

minimal  $\alpha$ -cap of  $F$  then  $t_0$  is the  $(1 - \alpha)$ -quantile of the variable  $u'_0 X$ , where  $X$  has distribution  $F$ .

The minimal  $\alpha$ -cap of a distribution  $F$  is unique with center  $u_0 \in S$  if and only if  $F$  satisfies the following property:

(A) *For every  $u \neq u_0$  and every  $t$  such that  $F(u, t) \geq \alpha$ , there exists  $t' > t$  such that  $F(u_0, t') \geq \alpha$ .*

We shall denote by  $\mathcal{M}_\alpha(u_0)$  the class of absolutely continuous distributions on  $S$  verifying property (A) and by  $\mathcal{M}(u_0)$  the class  $\bigcap_\alpha \mathcal{M}_\alpha(u_0)$ . An absolutely continuous distribution  $F$  belongs to the latter class if and only if it satisfies the following “unimodality” property:

(B) *For every  $u \neq u_0$  and every  $t$ , there exists  $t' > t$  such that  $F(u_0, t') \geq F(u, t)$ .*

An absolutely continuous distribution with unimodal density does generally not satisfy property (B). It is however the case if moreover  $F$  is rotationally symmetric about the mode of its density. The Langevin distribution is such an example. Its density is proportional to  $\exp(ku'_0 x)$ , and so is both unimodal at  $u_0$  and rotationally symmetric about  $u_0$ . Not surprisingly, the  $\alpha$ -concentration coefficient is a strictly increasing function of the concentration parameter  $k$  appearing in the density. For  $p = 3$  we have  $t_0 = k^{-1} \log[e^k - 2\alpha \sinh(k)]$ . On the other hand, if  $f$  is rotationally symmetric and bimodal, then there are two minimal  $\alpha$ -caps  $C(u_0, t_0)$  and  $C(-u_0, t_0)$ , which provide a unique axis of rotation. Finally, the normalized mean (see Watson, 1983) and the normalized median (see Ducharme and Milasevic, 1987) both coincide with  $u_0$  if  $F$  belongs to  $\mathcal{M}(u_0)$ .

The following algorithm will generally allow us to find a minimal  $\alpha$ -cap of a set  $X^{(n)}$  of points on  $S$ . It is stated for the case  $p = 3$  but can be generalized to any dimension. It points out the similarity between the minimal cap and the minimum-volume sphere.

**S1.** *For each triple of points of  $X^{(n)}$  consider the circle on  $S^2$  determined by them. Sort the circles according to the number of elements of  $X^{(n)}$  in them, and for each  $[\alpha n]$ , choose among the caps with  $[\alpha n]$  elements those with minimal opening.*

**S2.** *For each pair of points of  $X^{(n)}$  consider the smallest circle on  $S^2$  containing the two points and then proceed as in S1.*

**S3.** *Let  $C_n^*$  be the smallest of the two caps obtained in S1 and S2.*

**Proposition 1** *If  $X^{(n)}$  is in general position, i.e. no more than  $p$  points of  $X^{(n)}$  lie on an affine hyperplane, then  $C_n^*$  is a minimal  $\alpha$ -cap of  $X^{(n)}$ .*

The proposition states that every minimal cap can be found via the above algorithm. This follows from the fact that the boundary of a minimal cap must contain two or three points of  $X^{(n)}$ . Otherwise, there would exist a smaller cap with the same number of elements, which contradicts the minimality property.

To reduce computation time, the algorithm proposed by Rousseeuw and Leroy (1987) can be adapted to this problem. Rather than searching over all caps determined by the  $\binom{n}{3} + \binom{n}{2}$  subsets of observations, a set of  $m$  triples of observations can be selected at random. For each triple, the corresponding cap is computed; then the cap is shrunk or expanded until it contains  $\lceil \alpha n \rceil$  points. The minimal cap is chosen from among these  $m$   $\alpha$ -caps, and the order of operations is now reduced to  $O(mn)$ .

### 3 Asymptotic properties

Let  $X^{(n)}$  be an i.i.d. sample of size  $n$  from  $F \in \mathcal{M}_\alpha(u_0)$ , and let  $C(u_n, t_n)$  be a minimal  $\alpha$ -cap of  $X^{(n)}$ . Note that  $X^{(n)}$  is almost surely in general position. We then have the following

**Proposition 2** *Assume that  $F(u_0, t_0 - \delta) > \alpha$  for each (admissible)  $\delta > 0$ . Then i)  $t_n \rightarrow t_0$  almost surely as  $n \rightarrow \infty$  and ii)  $u_n \rightarrow u_0$  almost surely as  $n \rightarrow \infty$ .*

In the case where  $F$  is rotationally symmetric about  $u_0$ , its density is of the form  $f(u'_0 x)$  for some suitable function on  $[-1, 1]$ . The assumption of the proposition is then satisfied if, for example,  $f$  is continuous at  $t_0$  and  $f(t_0) > 0$ .

Note that  $u_n$  can be defined as any unit vector maximizing the  $(1 - \alpha)$ -quantile of the set  $\{u'X_1, \dots, u'X_n\}$  and that  $t_n$  is the value of this maximal quantile. We shall prove that the asymptotic distribution of  $t_n$  is the same as that of the empirical  $(1 - \alpha)$ -quantile of the variable  $u'_0 X$ . More precisely, the following result holds.

**Proposition 3** *Assume that  $F \in \bigcap_{|\beta - \alpha| < \epsilon} \mathcal{M}_\beta(u_0)$  for some  $\epsilon > 0$  and that the density  $p_0$  of  $u'_0 X$  is positive and continuous at  $t_0$ . Then*

$$\sqrt{n}(t_n - t_0) \rightarrow \mathcal{N}\left(0, \frac{\alpha(1 - \alpha)}{p_0^2(t_0)}\right).$$

Now turn to the behavior of  $u_n$ . For the next result, we assume that  $F \in \mathcal{M}_\alpha(u_0)$  and that  $F$  has a rotationally symmetric density of the form  $f(u'_0 x)$  for some  $f$  defined on  $[-1, 1]$ . Moreover, we only consider the cases where  $p \geq 3$  and  $\alpha \leq 1/2$ . The case  $p = 2$  amounts to the one-dimensional context



of the shorth, which is treated in Kim and Pollard (1990). The restriction on  $\alpha$  ensures  $t_0 \geq 0$ , and greatly simplifies the covariance structure of the limit process.

Denote by  $I_q$  the  $q$ -dimensional identity matrix and by  $B_{p-1}$  the unit closed ball  $\{u^* \in R^{p-1}; |u^*| \leq 1\}$ . For a measurable subset  $E \subset S^q$ ,  $\text{Area}(E)$  represents the  $q$ -dimensional volume of  $E$ . Given a function  $h$  and a distribution  $P$ , write  $Ph$  for the expectation of  $h$  under  $P$ . Let  $Z(x)$  be a Gaussian process indexed by  $R^{p-1}$  with continuous sample paths, zero expectation and covariance kernel

$$\Gamma(x, y) = (p-2)^{-1} \text{Area}(S^{p-3}) f(t_0)(1-t_0^2)^{\frac{p-2}{2}} (|x| + |y| - |x-y|).$$

**Proposition 4** *Assume that  $f$  is twice differentiable on  $(-1, 1)$  and that  $f'(t_0) > 0$ . Then*

$$n^{1/3}(I_p - u_0 u'_0)u_n \xrightarrow{\mathcal{L}} x_{\max},$$

where  $x_{\max}$  is the almost surely unique vector maximizing  $Z(x) + \frac{1}{2}x'Ux$ , with

$$U = -\frac{1}{p-1} \text{Area}(S^{p-2}) f'(t_0)(1-t_0^2)^{\frac{p-1}{2}} I_{p-1}.$$

**Corollary 1** *Under the hypotheses of Proposition 4,*

$$2n^{2/3}(1 - u'_0 u_n) \xrightarrow{\mathcal{L}} |x_{\max}|^2.$$

The two different rates of convergence, cube-root for the direction and square-root for the concentration, parallel those of the center and length of the shorth. The minimal caps can be extended to the excess-mass approach by finding the cap that maximizes

$$F_n(u, t) - \alpha \text{Area}(C(u, t)).$$

In this case, both the direction and opening of the excess-mass cap would have cube-root rates of convergence. Properties of the excess-mass cap are not addressed here.

## 4 Proofs

### Proof of Proposition 2

i) The class of sets  $\mathcal{C} = \{C(u, t); u \in S, -1 \leq t \leq 1\}$  is a Vapnik-Cervonenkis class and thus (e.g. Pollard, 1984)

$$\gamma_n = \sup_{u, t} |F_n(u, t) - F(u, t)| \xrightarrow{a.s.} 0.$$

It follows that for each  $\delta > 0$ ,  $F_n(u_0, t_0 - \delta) > \alpha$  eventually, almost surely. This implies that, for  $n$  large enough,  $t_n \geq t_0 - \delta$  except on a set of measure zero. Moreover, the almost sure convergence of  $\gamma_n$  to 0 implies that

$$\sup_u |F_n(u, t_0 + \delta) - F(u, t_0 + \delta)| \xrightarrow{a.s.} 0,$$

which in turn gives the almost sure upper bound  $t_n < t_0 + \delta$  for  $n$  large enough.

ii) Note first that for each  $\eta > 0$  there exists  $\epsilon > 0$  such that

$$\sup_{\|u-u_0\| \geq \eta} F(u, t_0) = \alpha - \epsilon.$$

It then follows from continuity of  $F(.,.)$  that there exists  $\delta > 0$  such that

$$\sup_{\|u-u_0\| \geq \eta, |t-t_0| \leq \delta} F(u, t) < \alpha - \epsilon/2.$$

Consequently, from i,

$$\sup_{\|u-u_0\| \geq \eta} F(u, t_n) < \alpha - \epsilon/2$$

eventually, almost surely. The latter inequality and  $\gamma_n \xrightarrow{a.s.} 0$  imply that  $\|u_n - u_0\| < \eta$ , almost surely.  $\square$

### Proof of Proposition 3

The idea is to wedge  $t_n$  between two random variables having the same asymptotic behavior. Let  $s_n$  be the  $(1 - \alpha)$  sample quantile of  $\{u'_0 X_1, \dots, u'_0 X_n\}$ . Then by definition

$$F_n(u_0, s_n) = \alpha + O_P(1/n)$$

and it follows from the minimality of  $C(u_n, t_n)$  that  $s_n \leq t_n$ .

The hypotheses of the proposition entail (e.g. see Serfling, 1980) that

$$\sqrt{n}(s_n - t_0) \longrightarrow \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{p_0^2(t_0)}\right).$$

Now we find an upper bound for  $\sqrt{n}(t_n - t_0)$  with the same asymptotic distribution as  $\sqrt{n}(s_n - t_0)$ . We have

$$\begin{aligned} \alpha + O_P(1/n) &= F_n(u_n, t_n) \\ &= F(u_n, t_n) + (F_n - F)(u_n, t_n) \\ &\leq F(u_0, t_n) + (F_n - F)(u_n, t_n) \\ &= \alpha - p_0(t_0)(t_n - t_0) + o_P(t_n - t_0) + (F_n - F)(u_0, t_0) \\ &\quad + o_P(n^{-1/2}). \end{aligned}$$

The inequality follows from the hypothesis on  $F$  and consistency of  $t_n$ . It holds on a set  $A_n$  of probability tending to one. Note here that the hypothesis of Proposition 2 is satisfied. The last  $o_P$  term follows from consistency of  $u_n$  and  $t_n$  and the fact that the process  $\sqrt{n}(F_n - F)$  indexed by the class  $\mathcal{C}$  is stochastically equicontinuous. This sequence of inequalities imply that on  $A_n$

$$\sqrt{n}(t_n - t_0) \leq \frac{1}{p_0(t_0)} \sqrt{n}(F_n - F)(u_0, t_0) + o_P(1).$$

Therefore, the random variable on the right hand side converges weakly to the desired normal distribution (Pollard, 1984, Theorem VII.21).  $\square$

### Proof of Proposition 4

We may assume without loss of generality that  $u_0 = (0, \dots, 0, 1)'$ . For  $u^* \in B_{p-1}$ , let  $u = u(u^*)$  be the point on  $S^{p-1}$  above  $u^*$ , i.e.  $u = (u^*, \sqrt{1 - |u^*|^2})'$ . For  $n$  large enough,  $u_n$  can be uniquely represented as  $u_n = u(u_n^*)$  for some  $u_n^* \in B_{p-1}$ . Note that the north pole  $u_0$  corresponds to  $u(0)$ . Next, define

$$W(., u^*, \delta) = C(u(u^*), t_0 + \delta) - C(u_0, t_0 + \delta)$$

which is to be understood as the difference of the indicator functions of the corresponding sets.

Note that  $u_n^*$  is a solution of the maximization  $\sup_{u^*} F_n W(., u^*, t_n - t_0)$ . To prove this proposition we shall use the main theorem of Kim and Pollard (1990). To do so, we need to establish two results. The first is that  $u_n$  also maximizes  $F_n W(., u^*, 0)$ , i.e.

$$F_n W(., u_n^*, 0) \geq \sup_{u^*} F_n W(., u^*, 0) - o_p(n^{-2/3}).$$

To show this we need the following lemma.

**Lemma 1** *Define the function  $M : B_{p-1} \times [-1, 1] \rightarrow [0, 1]$  by  $M(u^*, t) = F(u, t)$  and let*

$$\gamma(t) = -\frac{1}{p-1} \text{Area}(S^{p-2}) f'(t) (1 - t^2)^{(p-1)/2}.$$

*Then we have the expansion*

$$\begin{aligned} M(u^*, t_0 + \delta) = & \alpha + \frac{\partial}{\partial t} \Big|_{t_0} M(0, t) \delta + \frac{1}{2} \frac{\partial^2}{\partial t^2} \Big|_{t_0} M(0, t) \delta^2 \\ & + \frac{1}{2} \gamma(t_0) |u^*|^2 + o(\delta^2) + o(|u^*|^2). \end{aligned}$$

**Proof:** We have  $M(u^*, t) = \int_{C(u_0, t)} f(u'_0 T_u x) dS(x)$  where  $T_u$  is the rotation mapping  $u_0$  to  $u$ . Note that  $T_{u_0} = I_p$ . Taking into account the fact that  $u_0$  is the north-pole, that the  $p$ -th row of the matrix  $T_u$  is  $(-u^*, \sqrt{1 - |u^*|^2})$ , and using the measure decomposition  $dS^{p-1} = (1 - s^2)^{(p-3)/2} ds \otimes dS^{p-2}$  (see e.g. Watson, 1983), we obtain after some calculations that for each  $t$  and each  $j, k \in \{1, \dots, p-1\}$ ,

$$\nabla|_{u^*=0} M(u^*, t) = 0 = \frac{\partial^2}{\partial u_j^* \partial u_k^*}|_{u^*=0} M(u^*, t)$$

and

$$\frac{\partial^2}{\partial u_j^{*2}}|_{u^*=0} M(u^*, t) = \gamma(t). \quad \square$$

As a consequence of Lemma 1,

$$FW(., u^*, \delta) = \frac{1}{2} \gamma(t_0) |u^*|^2 + o(\delta^2) + o(|u^*|^2).$$

Use this equation, the fact that  $\gamma(t_0) < 0$ , and stochastic equicontinuity of  $n^{2/3}(F_n - F)$  (which follows Kim and Pollard) in order to obtain the inequality

$$F_n W(., u_n^*, 0) \geq \sup_{u^*} F_n W(., u^*, 0) - o_p(n^{-2/3}).$$

The second result that needs to be established is that the limiting covariance is, for each  $x, y \in R^{p-1}$ ,

$$\Gamma(x, y) = \lim_{\beta \downarrow 0} \beta^{-1} FW(., \beta x, 0) W(., \beta y, 0).$$

Let  $t \geq 0$  and, given two vectors  $u, v \in S$ , consider the set

$$J_t(u, v) = C(u, t) \setminus (C(u, t) \cap C(v, t)).$$

Write  $A_t(\xi)$  for its area, where  $\xi$  denotes the angle between  $u$  and  $v$ . To obtain this result, we need the following lemma, which has interest in its own right.

**Lemma 2** *The value of the right hand derivative of  $A_t$  at  $\xi = 0$  is given by*

$$A'_t(0) = (p-2)^{-1} \text{Area}(S^{p-3}) (1 - t^2)^{\frac{p-2}{2}}.$$

**Proof:** We shall compute the right hand derivative of  $B_t(\xi) = \text{Area}(C(u, t) \cap C(v, t))$ . Note that  $A'_t(0) = -B'_t(0)$ . The cap  $C(u, t)$  can be viewed as

a  $(p-1)$ -dimensional spherical disc on  $S$  with radius  $r = \arccos(t)$ . Its boundary is a  $(p-2)$ -dimensional sphere on  $S$  with center  $u \in S$  and (spherical) radius  $r$ . We shall denote it by  $S^{p-2}(u, r)$  and we shall denote by  $S^{p-2}(u, r, \phi)$  a cap on  $S^{p-2}(u, r)$  with half-angle  $\phi$  (there is no need for our purposes to specify its center on  $S^{p-2}(u, r)$ ). We have

$$\text{Area}(S^{p-2}(u, r, \phi)) = \sin^{p-2}(r) \text{Area}(S^{p-3}) \int_0^\phi \sin^{p-3}(\theta) d\theta.$$

Now let  $S^+$  be the hemisphere of  $S$  determined by  $S^{p-2}(u, r) \cap S^{p-2}(v, r)$  which contains  $v$ . The Riemannian distance from  $u$  to  $S^+$  is given by  $a = \xi/2$ . Denote by  $L^{p-1}(a, r)$  the intersection of  $C(u, t)$  and  $S^+$ . Then  $B_t(\xi) = 2 \text{Area}(L^{p-1}(a, r))$ .

Now,  $L^{p-1}(a, r)$  is composed of caps of the form  $S^{p-2}(u, r, \phi(\rho))$ , where  $a \leq \rho \leq r$  and where  $\phi(\rho)$  is given by the spherical trigonometry formula

$$\tan(\phi(\rho)) = \frac{\sqrt{\cos^2(a) - \cos^2(\rho)}}{\sin(a) \cos(\rho)}.$$

It follows that

$$B_t(\xi) = 2 \text{Area}(S^{p-3}) \int_a^r \sin^{p-2}(\rho) \left( \int_0^{\phi(\rho)} \sin^{p-3}(\theta) d\theta \right) d\rho$$

and we obtain after some computations that

$$B'_t(0) = -\frac{1}{p-2} \text{Area}(S^{p-3}) \sin^{p-2}(r). \quad \square$$

Now, for  $\beta$  small enough, denote by  $x_\beta, y_\beta$  the vectors on  $S^{p-1}$  which are above  $\beta x, \beta y$ . Then, from continuity of  $f$  and Lemma 2, we obtain

$$\begin{aligned} & \lim_{\beta \downarrow 0} \beta^{-1} FW(., \beta x, 0) W(., \beta y, 0) \\ &= \lim_{\beta \downarrow 0} \beta^{-1} (F(J_{t_0}(x_\beta, u_0)) + F(J_{t_0}(u_0, y_\beta)) - F(J_{t_0}(x_\beta, y_\beta))) \\ &= f(t_0) A'_{t_0}(0) \lim_{\beta \downarrow 0} \beta^{-1} [\xi(x_\beta, u_0) + \xi(u_0, y_\beta) - \xi(x_\beta, y_\beta)] \\ &= \Gamma(x, y). \quad \square \end{aligned}$$

## References

- [1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W. (1972). *Robust Estimation of Location: Survey and Advances*. Princeton University Press.

- [2] Arcones, M.A., Chen, Z., and Gine, E. (1994). Estimators related to U-processes with applications to multivariate medians: Asymptotic normality. *Ann. Statist* **22**, 1460-1477.
- [3] Brown, B.M. (1983). Statistical uses of the spatial median. *J.R. Statist. Soc. B* **45** 25-35.
- [4] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J.R. Statist. Soc* **91**, 862-872.
- [5] Chernoff, H. (1964). Estimation of the mode. *Ann. Math. Statist* **16**, 31-41.
- [6] Donoho, D.L., and Gasko, M. (1992) Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20**, 1803-1827.
- [7] Ducharme, G.R., Milasevic, P. (1987). Spatial median and directional data. *Biometrika* **74**, 212-215.
- [8] Dümbgen, L. (1992) Limit theorems of the simplicial depth. *Statist. Prob. Lett.* **14**, 119-128.
- [9] Eddy, W.F., and Hartigan, J.A. (1977). Uniform convergence of the empirical distribution function over convex sets. *Ann. Statist.* **5**, 370-374.
- [10] Einmahl, J.H.J., and Mason, D.M. (1992). Generalized quantile processes. *Ann. Statist.* **20**, 1062-1078.
- [11] Fisher, N.I., Lewis, T., Embleton, B.J.J. (1987). *Statistical Analysis of Spherical Data*. Cambridge University Press.
- [12] Grübel, R. (1988). The length of the shorth. *Ann. Statist.* **16**, 619-629.
- [13] Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *J. Am. Statist. Soc.* **82**, 267-270.
- [14] Kim, J., and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191-219.
- [15] Liu, R.Y. (1990). On a notion of data depth based upon random simplices. *Ann. Statist.* **18**, 405-414.
- [16] Liu, R.Y., and Singh, K. (1992) Ordering directional data: Concepts of data depth on circles and spheres. *Ann. Statist.* **20**, 1468-1484.
- [17] Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests of multimodality. *J. Am. Statist. Soc.* **86**, 738-746.
- [18] Nolan, D. (1991). The excess-mass ellipsoid. *J. Multiv. Anal.* **39**, 348-371.
- [19] Nolan, D. (1992). Asymptotics for multivariate trimming. *Stoch. Proc. Appl.* **42**, 157-169.
- [20] Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.
- [21] Polonik, W. (1995a). Measuring mass concentrations and estimating

- density contour clusters - An excess mass approach. *Ann. Statist.* **23**, 855-881.
- [22] Polonik, W. (1995b). Density estimation under qualitative assumptions in higher dimensions. *J. Multiv. Anal.* **55**, 61-81.
- [23] Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
- [24] Sager, T.W. (1979). An iterative method for estimating a multivariate mode and isopleth. *J. Am. Statist. Soc.* **74** 329-339.
- [25] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
- [26] Small, C.G. (1990) A survey of multidimensional medians. *Int. Statist. Rev.* **58**, 263-277.
- [27] Venter, J.H. (1967). On estimation of the mode. *Ann. Math. Statist.* **38**, 1446-1455.
- [28] Watson, G.S.(1983). *Statistics on Spheres*. New York: Wiley.

# An $L_1$ -norm procedure for fitting overlapping clustering models to proximity data

Anil Chaturvedi

*AT&T Laboratories, USA*

J. Douglas Carroll

*Rutgers University, New Jersey, USA*

*Abstract:* We describe a general approach to finding Least Absolute Deviation estimates of two-way and three-way overlapping clustering models called ADCLUS and INDCLUS. The suggested approach utilizes a combinatorial optimization approach that takes advantage of a separability property of this loss function for fitting these models. Our approach helps in robustifying the solutions in the presence of extreme outliers in the data.

*Key words:* Cluster analysis, overlapping clusters, combinatorics, trilinear models, least absolute deviation.

AMS subject classification: Primary 62H30; secondary 92J45.

## 1 Introduction

Shepard and Arabie (1979) first introduced an overlapping clustering model for classification based on similarity data called ADCLUS (for additive clustering). Subsequently, Arabie and Carroll (1980), provided a mathematical programming approach for fitting this model. Carroll and Arabie (1983), proposed an individual differences generalization of this model, which they called INDCLUS (for individual differences clustering), and also devised a procedure for fitting this three-way generalization, thus providing a methodology for three-way overlapping clustering. These algorithms optimize a least squares, or  $L_2$ -norm based, loss function. The theoretic-



cal significance and conceptual elegance of overlapping clustering can be seen in applications in many different substantive domains, as illustrated, for example, in Arabie, Carroll, DeSarbo, and Wind (1981) and Srivastava, Alpert, and Shocker (1984). In this paper, we present a procedure for fitting these models via an  $L_1$ -norm, which we call LADCLUS (for Least Absolute Deviation clustering). This procedure is a special case of a general approach (Carroll and Chaturvedi, 1995) for fitting a general multilinear model including both discrete and continuous parameters via  $L_1$ - or  $L_2$ -norms, or other  $L_p$ -norms. Our procedure is computationally significantly faster and can handle much larger data sets than Lakshmi-Ratan's (1985)  $L_1$ -norm based approach for fitting the two-way overlapping clustering problem. It is computationally as simple and tractable as a procedure proposed by Chaturvedi and Carroll (1994), which provides a more efficient algorithm than the earlier algorithms for fitting these models via an  $L_2$ -norm. The principal benefit of fitting these overlapping clustering models via an  $L_1$ -norm would be to robustify the estimation of model parameters vis-a-vis extreme outliers in the data. Fitting models via an  $L_1$ -norm tends to reduce the effects of extreme outliers, as compared to fitting via an  $L_2$  (OLS)-norm (Hampel, Ronchetti, Rousseeuw and Stahel, 1986; Kaufman and Rousseeuw, 1990). The increased robustness of the  $L_1$ -norm procedure to the  $L_2$ -norm procedure is analogous to the illustration in Rousseeuw and Leroy (1987, pp. 10-11) in the context of linear regression, wherein the  $L_1$ -norm estimate is shown to be more robust to extreme values of the dependent variable. The derived solutions would be more robust to extreme values in the data (analogous to the dependent variable in the regression case).

## 2 The INDCLUS model

Assume that  $N$  objects are being clustered into  $R$  overlapping clusters. Then, the INDCLUS model (Carroll, 1975; Carroll and Arabie, 1983) is written as:

$$\mathbf{S}_k = \mathbf{P}\mathbf{W}_k\mathbf{P}' + \mathbf{C}_k + \text{error}, \quad (1)$$

where:

$\mathbf{S}_k$  is an  $(N \times N)$  similarity matrix for the  $k$ th subject (or other source of data);  $k = 1, \dots, K$ ,

$\mathbf{W}_k$  is an  $(R \times R)$  diagonal matrix of weights for the  $k$ th subject (or other source of data);  $k = 1, \dots, K$ ,

$\mathbf{P}$  is an  $(N \times R)$  *binary* indicator matrix defining the possibly overlapping clusters, and

$\mathbf{C}_k$  is an  $(N \times N)$  matrix, all of whose entries are  $c_k$ , which can be thought of as the weight for a universal cluster denoted by an  $N \times 1$  unit vector  $\mathbf{1}$ , all of whose components are 1.

The diagonal entries in the  $(N \times N)$  matrix  $\mathbf{S}_k$  are usually not defined. The estimation problem is to determine the ordinary least squares (OLS) estimates of parameters  $\mathbf{P}$ ,  $\mathbf{W}_k$  and  $\mathbf{C}_k$ . The diagonal elements of the weight matrices  $\mathbf{W}_k$  must be non-negative and elements of  $\mathbf{P}$  must be constrained to be either 0 or 1. The ADCLUS model is the special case of the INDCLUS model in which  $K = 1$ . The ADCLUS procedure proposed by Shepard and Arabie (1979) combines a combinatorial algorithm with iterative estimation of the weights. The MAPCLUS procedure proposed by Arabie and Carroll (1980) fits the ADCLUS model via a penalty function based mathematical programming technique embedded in an overall alternating least squares procedure. The INDCLUS method proposed by Carroll and Arabie (1983) generalizes the MAPCLUS approach to the three-way INDCLUS model. These are all OLS procedures (using an  $L_2$ -norm based fit measure). Various other techniques have been developed for fitting these models, such as the Maximum Likelihood approach of Hiroshi Hojo (1983), the Qualitative Factor Analysis (QFA) procedure of Mirkin (1987), and the OLS procedure called SINDCLUS, of Chaturvedi and Carroll (1992, 1994). We now describe the LADCLUS algorithm for fitting these models via an  $L_1$ -norm.

### 3 The LADCLUS algorithm

The LADCLUS procedure uses a separability property of the LAD loss function for fitting the INDCLUS model defined in (1) via an alternating estimation procedure discussed below. This separability property has also been used in the SINDCLUS procedure (Chaturvedi and Carroll, 1994). While Mirkin (1990) also uses a one-cluster-at-a-time approach to estimating a bilinear model, his approach utilizes a different algorithm for estimating the parameters for a cluster. The two main differences of the approach presented in this paper with Mirkin's (1990) approach are: (a) Mirkin does not use the separability property, but a different approach to estimating the parameters for any given cluster, and (b) his approach does not yield overall  $L_1$ -norm based estimates for the model parameters, since he does not iterate over clusters (as we do) in order to obtain a globally optimal solution. This separability property can best be stated in the form of two procedures - the elementary discrete and elementary continuous LAD procedures. These are illustrated below.

### A. The elementary discrete LAD procedure

Consider the following illustrative problem. Let

$$\mathbf{L} = \begin{bmatrix} 7 & 5 & 3 & 9 \\ 8 & 6 & 5 & 1 \\ 9 & 4 & 2 & 7 \\ 5 & 3 & 4 & 6 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \quad \mathbf{r}' = \begin{bmatrix} 1 & 4 & 9 & 3 \end{bmatrix}$$

The estimation problem is to find the least absolute deviation (LAD) estimate of  $\mathbf{x}$ , where  $\mathbf{L} = \mathbf{x}\mathbf{r}' + \text{error}$  and  $\mathbf{x}$  is constrained to be binary (0 or 1). That is

$$\begin{bmatrix} 7 & 5 & 3 & 9 \\ 8 & 6 & 5 & 1 \\ 9 & 4 & 2 & 7 \\ 5 & 3 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 1x_1 & 4x_1 & 9x_1 & 3x_1 \\ 1x_2 & 4x_2 & 9x_2 & 3x_2 \\ 1x_3 & 4x_3 & 9x_3 & 3x_3 \\ 1x_4 & 4x_4 & 9x_4 & 3x_4 \end{bmatrix} + \text{error}$$

If we let

$$\begin{aligned} f_1 &= |7 - 1x_1| + |5 - 4x_1| + |3 - 9x_1| + |9 - 3x_1|, \\ f_2 &= |8 - 1x_2| + |6 - 4x_2| + |5 - 9x_2| + |1 - 3x_2|, \\ f_3 &= |9 - 1x_3| + |4 - 4x_3| + |2 - 9x_3| + |7 - 3x_3|, \text{ and} \\ f_4 &= |5 - 1x_4| + |3 - 4x_4| + |4 - 9x_4| + |6 - 3x_4|, \end{aligned}$$

then the sum of absolute errors is given by

$$F = f_1 + f_2 + f_3 + f_4.$$

Note that  $f_1$  is a function only of  $x_1$ ;  $f_2$  is a function only of  $x_2$ ;  $f_3$  is a function only of  $x_3$ ; and  $f_4$  is a function only of  $x_4$ . Thus,  $F$  is *separable* in  $x_1, x_2, x_3$ , and  $x_4$ . To minimize  $F$ , one can separately minimize  $f_1$  w.r.t  $x_1$ ,  $f_2$  w.r.t  $x_2$ ,  $f_3$  w.r.t  $x_3$ , and,  $f_4$  w.r.t.  $x_4$ . To minimize, say,  $f_1$  w.r.t.  $x_1$ , one can easily evaluate  $f_1$  at  $x_1 = 1$  and  $x_1 = 0$ . The  $x_1$  yielding a minimum of these two possible values is then chosen. Thus, for  $I$  (0-1) variables, only  $2I$  function evaluations and comparisons are needed, as compared to  $2^I$  evaluations and comparisons for explicit enumeration.

### B. The elementary continuous LAD procedure

Again, consider the illustrative problem given in the Elementary discrete LAD procedure of determining  $\mathbf{x}$ , where  $\mathbf{x}$  is real. As in the case of the Elementary discrete LAD procedure, it can be shown that  $F$  is *separable*

in  $x_1, x_2, x_3$ , and  $x_4$ . Thus, to minimize  $F$ , one can separately minimize  $f_1$  w.r.t  $x_1$ ,  $f_2$  w.r.t  $x_2$ ,  $f_3$  w.r.t  $x_3$ , and,  $f_4$  w.r.t.  $x_4$ . The minimization of, say,  $f_1$  w.r.t.  $x_1$ , is equivalent to performing simple  $L_1$  regression with a single independent variable. While the general multivariate  $L_1$ -regression problem can be formulated and solved as a constrained linear programming problem (Rousseeuw and Leroy 1987 pp 146), in our case since there is only one independent variable, the simplex solutions can easily be determined by evaluating the functions  $f_1 - f_4$  at the respective corner points. To illustrate, in order to minimize, say,  $f_1$  w.r.t.  $x_1$ , simply evaluate  $f_1$  at the four corner points given by  $x_1 = 7/1, 5/4, 3/9$  and  $9/3$ . The value of  $x_1$  yielding the minimum is chosen as the optimum LAD estimate of  $x_1$ . In this specific case,  $f_1$  is a minimum at  $x_1 = 3/9$ . More generally, to minimize the LAD criterion as a function of a single variable  $x_i$ , for the component  $f_i$  of the overall loss function

$$F = \sum_{i=1}^I f_i$$

where  $f_i = \sum_{j=1}^J |l_{ij} - x_i r_j|$  is a function only of  $x_i$ , we simply evaluate  $f_i$  at the  $J$  values

$$x_i^{(j')} = \frac{l_{ij'}}{r_{j'}}$$

and then choose the  $\hat{x}_i$  to be the  $x_i^{(j')}$  minimizing

$$f_i^{(j')} = \sum_{j=1}^J |l_{ij} - x_i^{(j')} r_j|.$$

This can easily be shown to provide the  $x_i$  minimizing  $f_i$ , thus completing the Elementary Continuous LAD procedure.

The LADCLUS procedure utilizes an alternating least absolute deviation ( $L_1$ -norm) procedure in fitting the ADCLUS/INDCLUS models. The algorithm converges to at least a local optimum, as the objective function value decreases (or does not increase) at each stage of the algorithm, and there is a lower bound to the objective function.

The estimation problem in LADCLUS is to find LAD estimates of  $\mathbf{P}$ ,  $\mathbf{W}_k$ , and  $\mathbf{C}_k$ , ( $k = 1, \dots, K$ ) in the equation:

$$\mathbf{S}_k = \mathbf{P}\mathbf{W}_k\mathbf{P}' + \mathbf{C}_k + error, \quad (2)$$

where  $\mathbf{P}$ ,  $\mathbf{W}_k$ , and  $\mathbf{C}_k$ , are as defined in equation (1). If  $\mathbf{S}_k$  were considered to be a nonsymmetric  $(N \times N)$  matrix, then the above equation can be generalized to

$$\mathbf{S}_k = \mathbf{P}\mathbf{W}_k\mathbf{Q}' + \mathbf{C}_k + error, \quad (3)$$

where  $\mathbf{Q}$  is an  $(N \times R)$  matrix, not necessarily the same as  $\mathbf{P}$ . In the *symmetric* case of INDCLUS,  $\mathbf{P} = \mathbf{Q}$ .

We use (3) to estimate the parameters for the symmetric case *without* imposing the constraint  $\mathbf{P} = \mathbf{Q}$ . Thus, we will first define the estimation problem for the nonsymmetric case, and then specialize it to the symmetric case. This is the same overall strategy as followed by Chaturvedi and Carroll (1994) in their SINDCLUS approach to OLS estimation of ADCLUS/INDCLUS. Defining the following symbols :

$\mathbf{p}_r = (N \times 1)$  binary vector for the  $r$ th cluster,

$\mathbf{w}_r = (K \times 1)$  vector of the weights for the  $r$ th cluster,

$\mathbf{q}_r = (N \times 1)$  binary vector for the  $r$ th cluster (not necessarily  $= \mathbf{p}_r$ ),

$\mathbf{P}_{(-r)} = (N \times R)$  binary matrix including the universal cluster  $\mathbf{1}$  but excluding the  $r$ th cluster,

$\mathbf{W}_{(-r)} = (R \times R)$  weight matrix for the  $k$ th subject or other source of data, including the weight for the universal cluster but excluding the weight for the  $r$ th cluster, and

$\mathbf{Q}_{(-r)} = (N \times R)$  matrix including the universal cluster  $\mathbf{1}$  but excluding the  $r$ th cluster,

we can rewrite (3) as

$$\mathbf{S}_k = \mathbf{p}_r w_{kr} \mathbf{q}'_r + \mathbf{P}_{(-r)} \mathbf{W}_{k(-r)} \mathbf{Q}'_{(-r)} + error,$$

If we have estimates of all but the  $r$ th cluster, then we can define  $\bar{\mathbf{S}}_k$  as

$$\bar{\mathbf{S}}_k = \mathbf{S}_k - \mathbf{P}_{(-r)} \mathbf{W}_{k(-r)} \mathbf{Q}'_{(-r)} \quad (4)$$

to get

$$\bar{\mathbf{S}}_k = \mathbf{p}_r w_{kr} \mathbf{q}'_r + error, \quad (5)$$

that is,

$$\bar{\mathbf{S}}_k = f(\text{parameters for cluster } r) + error \quad (6)$$

If we have  $K$  matrices  $\mathbf{S}_k$  of order  $(N \times N)$ , we can use a procedure similar to the CANDECOMP based algorithm called the INDSCAL method for fitting the INDSCAL model formulated by Carroll and Chang (1970). We

call this general procedure CANDCLUS (Carroll and Chaturvedi, 1995), which stands for CANonical Decomposition CLUStering. Let us assume that we have the parameter estimates for all clusters, except the  $i$ th cluster. Let  $\mathbf{T}_1$  be a  $(K \times N^2)$  matrix where the  $k$ th row has all the  $N^2$  terms of the  $N \times N$  matrix  $\bar{\mathbf{S}}_k$ , and  $\mathbf{T}_2$  and  $\mathbf{T}_3$  be  $(N \times KN)$  matrices).  $\mathbf{T}_1$  has all  $N^2$  elements of  $\bar{\mathbf{S}}_k$  in its  $k$ th row.  $\mathbf{T}_2$  is the supermatrix that has the  $j$ th row of matrices  $\bar{\mathbf{S}}_1, \dots, \bar{\mathbf{S}}_K$  in the  $j$ th row. Thus:

$$\mathbf{T}_2 = [\bar{\mathbf{S}}_1 \mid \bar{\mathbf{S}}_2 \mid \dots \mid \bar{\mathbf{S}}_k \mid \dots \mid \bar{\mathbf{S}}_K]$$

Similarly,

$$\mathbf{T}_3 = [\bar{\mathbf{S}}'_1 \mid \bar{\mathbf{S}}'_2 \mid \dots \mid \bar{\mathbf{S}}'_k \mid \dots \mid \bar{\mathbf{S}}'_K].$$

Assuming that estimates of  $\mathbf{p}_r$  and  $\mathbf{q}_r$  are known, the parameters for the  $r$ th cluster are estimated by iterating the following 3 Steps until *at least a local optimum* is reached.

• **Step 1. Estimating  $\mathbf{w}_r$  conditionally**

Given current estimates,  $\hat{\mathbf{p}}_r$  and  $\hat{\mathbf{q}}_r$ , of  $\mathbf{p}_r$  and  $\mathbf{q}_r$ , let  $\mathbf{g}_r$  be a vector of  $N^2$  elements such that

$$\mathbf{g}_r = \hat{\mathbf{p}}_r \otimes \hat{\mathbf{q}}_r,$$

where  $\otimes$  is the Kronecker product. Then, using the elementary continuous LAD procedure outlined earlier, one can find the LAD estimate  $\hat{\mathbf{w}}_r$  in the Equation

$$\mathbf{T}_1 = \mathbf{w}_r \mathbf{g}'_r + \text{error}.$$

Non-negativity constraints are imposed easily by simply setting all negative weights to zero as in Carroll, DeSoete, and Pruzansky (1989).

• **Step 2. Estimating  $\mathbf{p}_r$  conditionally**

Given current estimates,  $\hat{\mathbf{w}}_r$  and  $\hat{\mathbf{q}}_r$ , of  $\mathbf{w}_r$  and  $\mathbf{q}_r$ , let  $\mathbf{h}_r$  be a vector of  $KN$  elements such that

$$\mathbf{h}_r = \hat{\mathbf{w}}_r \otimes \hat{\mathbf{q}}_r.$$

Then, by using the elementary discrete LAD procedure in the equation

$$\mathbf{T}_2 = \mathbf{p}_r \mathbf{h}'_r + \text{error},$$

one can find  $\hat{\mathbf{p}}_r$ , the LAD estimates of  $\mathbf{p}_r$ .

• **Step 3. Estimating  $\mathbf{q}_r$  conditionally**

Given current estimates,  $\hat{\mathbf{w}}_r$  and  $\hat{\mathbf{p}}_r$ , of  $\mathbf{w}_r$  and  $\mathbf{p}_r$ , let  $\mathbf{j}_r$  be a vector of  $KN$  elements such that

$$\mathbf{j}_r = \hat{\mathbf{w}}_r \otimes \hat{\mathbf{p}}_r.$$

Then, by using the elementary discrete LAD procedure in the equation

$$\mathbf{T}_3 = \mathbf{q}_r \mathbf{j}_r' + \text{error},$$

one can find  $\hat{\mathbf{q}}_r$ , the LAD estimates of  $\mathbf{q}_r$ .

The LADCLUS algorithm starts off with random binary starting values for the matrices  $\mathbf{P}$  and  $\mathbf{Q}$ , and random continuous values in the diagonal matrices  $\mathbf{W}_k$ . The LADCLUS procedure iterates up to a prespecified maximum number of major iterations or until the percentage reduction in the fit value is less than a prespecified criterion. The current default is 0.0001. The current fit value (M2) and the previous fit value (M1) are updated after each major iteration. Each major iteration of the LADCLUS algorithm involves determining at least locally optimal conditional  $L_1$ -norm estimates of the parameters for the  $r+1$  clusters, using the one-cluster-at-a-time strategy. Thus, each major iteration comprises  $r+1$  minor iterations, corresponding to the  $r+1$  clusters.

The minor iteration of the LADCLUS procedure corresponding to the  $r^{th}$  cluster involves sequentially finding conditionally optimal parameter estimates  $\mathbf{w}_r$ ,  $\mathbf{p}_r$ , and  $\mathbf{q}_r$ . This is achieved by first forming supermatrices  $\mathbf{T}_1, \mathbf{T}_2$  and  $\mathbf{T}_3$ . The matrix  $\mathbf{g}_r$  is then formed and  $\hat{\mathbf{w}}_r$  is estimated as described in Step 1 using the elementary continuous LAD procedure. This is followed by Steps 2 and 3 of estimating  $\hat{\mathbf{p}}_r$  and  $\hat{\mathbf{q}}_r$  respectively, using the elementary discrete procedure. The fit value is then computed (F2), and compared to the previous fit value (F1). This process is repeated until there is no improvement in fit (i.e.  $F2 = F1$ ). At this point an (at least locally) optimal set of parameter estimates have been obtained for the  $r^{th}$  cluster, conditional on the fixed values of the  $R-1$  other clusters (and the universal cluster).

It should be noted that in Steps 2 and 3 above, the  $\hat{\mathbf{p}}$  and  $\hat{\mathbf{q}}$  vectors cannot be all zero. Thus, each cluster must have at least one object in it. Since the matrices  $\mathbf{S}_k$  usually do not have diagonals in the case of the INDCLUS model, Steps 1, 2 and 3 above need to be modified. In the case of undefined diagonals, we simply drop the corresponding columns from the  $\mathbf{T}_1$  matrix and  $\mathbf{g}$  vector in Step 1. Similarly, we don't consider the diagonal elements in Steps 2 and 3 for fitting the INDCLUS model. For estimating

the weights for the universal cluster, where  $\mathbf{p}$  and  $\mathbf{q}$  are fixed, we just use Step 1, with  $\mathbf{p} = \mathbf{q} = \mathbf{1}$ .

While we do not impose the constraint that  $\mathbf{P} = \mathbf{Q}$  in the estimation procedure, we find empirically that in the symmetric case of INDCLUS, the matrices  $\mathbf{P}$  and  $\mathbf{Q}$  are equal upon convergence at the global optimum. We conjecture that this will always occur under conditions of unique global optimality. Also, while the default option in LADCLUS is to estimate the INDCLUS model by treating the diagonals as missing data, LADCLUS also allows fitting of the INDCLUS model when diagonals are treated as non-missing. One major advantage of LADCLUS in addition to its greater computational efficiency, is its ability to handle arbitrary patterns of missing data satisfactorily. (In fact, the default option for diagonals simply is a special case of handling missing data). Since, at each stage of the algorithm, we are conditionally estimating one new “dimension” (e.g., the  $\mathbf{p}_r$  or  $\mathbf{q}_r$  vector in LADCLUS), through the use of the elementary discrete LAD procedure mentioned earlier, omission of data is accomplished by omitting the corresponding terms from the corresponding LAD loss function. (The treatment of missing data as described is a special case of weighted LAD fitting, with weights of zero for missing observations and one for those that are present.) The generalization of LADCLUS to *weighted* LAD, is also straightforward, since each of the three conditional LAD estimation stages can simply be replaced with an appropriately weighted LAD estimation procedure.

One final comment vis-a-vis LADCLUS is that no special case needs to be described for the fitting of the ADCLUS model, in which  $K = 1$ . ADCLUS is simply fit as a special case of the INDCLUS model, in which the third way, for subjects or other sources of data, has only one level.

## 4 Applications of LADCLUS to some real data

The LADCLUS procedure was applied to the Kinship data of Rosenberg and Kim published in Arabie, Carroll, and DeSarbo (1987). The application of SINDCLUS to this data set is described in detail in Chaturvedi and Carroll (1994). We present this application of LADCLUS to compare the solutions derived via LADCLUS and SINDCLUS.

The fifteen most commonly used kinship terms - Aunt, Brother, Cousin, Daughter, Father, Granddaughter, Grandfather, Grandmother, Grandson, Mother, Nephew, Niece, Sister, Son, and Uncle, were printed on slips of paper for use in a sorting task by Rosenberg and Kim (1975). Eighty-five male and eighty-five female subjects were run in a condition where subjects gave (only) a single-sort of the fifteen terms. A different group



of subjects (eighty-five males and eighty-five females) were told that, after making their first sorts of the terms, they should give additional subjective partitioning(s) of these stimuli using “a different basis of meaning each time”. Rosenberg and Kim (1975) used only the data from the first and second sortings for this group of subjects. Thus, we have six *conditions* which will correspond to our subjects: females’ single-sort, males’ single-sort, females’ first-sort, males’ first-sort, females’ second-sort, and males’ second-sort. Again note that the “subjects” (or other sources of data) in the first two conditions were distinct from those in the last four conditions.

Since the subjects’ partitions of the stimuli comprise nominal scale data that do not immediately assume the form of a proximity matrix, some pre-processing is necessary to obtain such a matrix. If we form a stimuli  $\times$  stimuli co-occurrence matrix for each experimental condition, with the  $(i, j)$ th entry derived as the number of subjects who placed stimuli  $i$  and  $j$  in the same group, and subtract that entry from the total number of subjects contributing to the matrix, then we have what is called the S-measure (Arabie, Carroll, and DeSarbo, 1987). As in Arabie, Carroll, and DeSarbo (1987), the six matrices constructed using the S-measure were analyzed using LADCLUS via a matrix unconditional approach. A five cluster solution explaining 38.75 percent of absolute deviation in the data (around the grand median) was extracted. The optimal clusters derived using the LADCLUS procedure are identical to the clusters derived by Arabie, Carroll, and DeSarbo (1987). The five cluster solution is presented in Table 1, while the importance weights derived via LADCLUS and INDCLUS are presented in Tables 2 and 3.

The clusters are easily interpreted. In the order listed, the first two are sex-defined, the third is the collateral relatives, the fourth is the nuclear family, while the fifth consists of grandparents and grandchildren.

## 5 Conclusions

The LADCLUS procedure introduced least absolute deviations as an objective function for the ADCLUS/INDCLUS models. The enhanced robustness of LADCLUS to the presence of extreme outliers over existing least squares procedures was demonstrated in an extensive Monte-Carlo simulation. The LADCLUS procedure can be extended to fit other hybrid multivariate models that entail both continuous and discrete parameters, where the discrete parameters can take on any discrete values. We hope that the LADCLUS procedure further enhances the applicability of the ADCLUS and INDCLUS models. While in the current version LADCLUS can yield locally optimal solutions, we are investigating approaches that

ameliorate these problems and yield globally optimal solutions, even for large data sets.

Table 1: LADCLUS & INDCLUS solutions for Rosenberg and Kim data.

| Cluster  | Items in Cluster                                                           | Interpretation                     |
|----------|----------------------------------------------------------------------------|------------------------------------|
| <i>a</i> | Brother, father,<br>grandfather<br>grandson, nephew,<br>son, uncle         | Male relatives excluding cousins   |
| <i>b</i> | Aunt, daughter,<br>granddaughter,<br>grandmother,<br>mother, niece, sister | Female relatives excluding cousins |
| <i>c</i> | Aunt, cousin,<br>nephew, niece, uncle                                      | Collateral relatives               |
| <i>d</i> | Brother, daughter,<br>father, mother,<br>sister, son                       | Nuclear family                     |
| <i>e</i> | Granddaughter,<br>grandfather,<br>grandmother,<br>grandson                 | Grandparents and Grandchildren     |
| <i>f</i> | All objects                                                                | Universal cluster                  |

Table 2: LADCLUS weights for the Rosenberg and Kim data.

| Subject   | <i>a</i> | <i>b</i> | <i>c</i> | <i>d</i> | <i>e</i> | Universal |
|-----------|----------|----------|----------|----------|----------|-----------|
| F' single | 0.11     | 0.11     | 0.44     | 0.29     | 0.48     | 0.04      |
| M' single | 0.2      | 0.19     | 0.29     | 0.28     | 0.28     | 0.05      |
| F' first  | 0.57     | 0.57     | 0.27     | 0.17     | 0.21     | 0.05      |
| F' second | 0.25     | 0.27     | 0.33     | 0.25     | 0.32     | 0.09      |
| M' first  | 0.33     | 0.33     | 0.28     | 0.17     | 0.29     | 0.08      |
| M' second | 0.31     | 0.32     | 0.11     | 0.15     | 0.13     | 0.16      |

Table 3: INDCLUS weights for the Rosenberg and Kim data.

| Subject   | <i>a</i> | <i>b</i> | <i>c</i> | <i>d</i> | <i>e</i> | Universal |
|-----------|----------|----------|----------|----------|----------|-----------|
| F' single | .052     | .049     | .552     | .478     | .626     | .055      |
| M' single | .143     | .146     | .397     | .372     | .449     | .075      |
| F' first  | .551     | .554     | .283     | .206     | .251     | .132      |
| F' second | .241     | .246     | .373     | .322     | .385     | .158      |
| M' first  | .299     | .291     | .340     | .241     | .395     | .158      |
| M' second | .295     | .306     | .237     | .219     | .253     | .207      |

References

[1] Arabie, P., and Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika* **45**, 211-235.

[2] Arabie, P., Carroll, J. D., and DeSarbo, W. S. (1987). Three-way Scaling and Clustering. Sage, Newbury Park, California.

[3] Arabie, P., Carroll, J. D., DeSarbo, W. S., and Wind, J. (1981). Overlapping clustering: a new method for product positioning. *Journal of Marketing Research* **18**, 310-17.

[4] Carroll, J. D., (1975). Models for individual differences in similarities. Paper presented at the Eighth Annual Mathematical Psychology Meeting, Purdue University, West Lafayette, IN.

[5] Carroll, J. D. and Arabie, P. (1983). An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika* **48**, 157-169.

[6] Carroll, J. D., and Chang J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, 283-319.

[7] Carroll, J. D., and Chaturvedi, A. D. (1995). A general approach to clustering and multidimensional scaling of two-way, three-way, or higher-way data. In *Geometric Representations of Perceptual Phenomena: Articles in Honor of Tarow Indow's 70th Birthday*, Eds. R. D. Luce, M. D'Zmura, D. D. Hoffman, G. Iverson, and A. K. Romney. In press.

[8] Carroll, J. D., DeSoete, G., and Pruzansky, S. (1989). Fitting of the latent class model via iteratively reweighted least squares CANDECOMP with nonnegativity constraints. In *Multiway Data Analysis*, Eds. R. Coppi and S. Bolasco, pp. 463-472. Amsterdam: North Holland.

[9] Carroll, J. D. and Pruzansky, S. (1975). Fitting of hierarchical tree

- structure (HTS) models, mixtures of HTS models, and hybrid models, via mathematical programming and alternating least squares. *Proc. of the U.S.-Japan Seminar on Multidimensional Scaling*, 9-19.
- [10] Carroll, J. D. and Pruzansky, S. (1980). Discrete and hybrid scaling models. In *Similarity and Choice*, Eds. E. D. Lantermann and H. Feger, pp. 108-139. Bern: Hans Huber.
  - [11] Chaturvedi, A. D. (1993). Perceived product uniqueness in product differentiation and product choice. *9316296*, University Microfilms International, Ann Arbor, Michigan.
  - [12] Chaturvedi, A., D., and Carroll, J. D. (1992). ALCLUS: alternating least squares clustering. Presented at the Annual Meeting of The Classification Society of North America, East Lansing, Michigan, June 11-13.
  - [13] Chaturvedi, A. D., and Carroll, J. D. (1994). An alternating combinatorial optimization approach to fitting the INDCLUS and generalized INDCLUS models. *Journal of Classification* **11**, 155-170.
  - [14] Chaturvedi, A. D., Carroll, J. D., and Lakshmi-Ratan, R. A. (1993). MADCLUS: minimum absolute deviation clustering. Presented at the Annual Meeting of The Classification Society of North America, Pittsburgh, Pennsylvania, June 24-26.
  - [15] DeSarbo, W. S. (1982). GENNCLUS: new models for general nonhierarchical clustering analysis. *Psychometrika* **47**, 449-475.
  - [16] Hartigan, J. A. (1975). *Clustering Algorithms*. New York: Wiley.
  - [17] Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior* **8**, 176-184.
  - [18] Hiroshi, Hojo (1983). A maximum likelihood method for additive clustering and its applications. *Japanese Psychological Research* **25**, 191-201.
  - [19] Johnson, S. C. (1967). Hierarchical Clustering Schemes. *Psychometrika* **32**, 241-254.
  - [20] Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data*. New York: Wiley Inter-Science.
  - [21] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.
  - [22] Lakshmi-Ratan, R. A. (1985). A linear mixed integer programming approach to overlapping cluster analysis. Working paper, University of Wisconsin, Madison.
  - [23] Mirkin, B. G. (1987). Additive clustering and qualitative factor analysis methods for similarity matrices. *Journal of Classification* **4**, 7-31.
  - [24] Mirkin, B. G. (1989). Erratum. *Journal of Classification* **6**, 271-272.

- [25] Mirkin, B. G. (1990). A sequential fitting procedure for linear data analysis models. *Journal of Classification* **7**, 167-195.
- [26] Rosenberg, S., and Kim, M. P. (1975). The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research* **10**, 489-502.
- [27] Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
- [28] Shepard, R. N. and Arabie, P. (1979). Additive clustering representation of similarities as combinations of discrete overlapping properties. *Psychological Review* **86**, 87-123.
- [29] Srivastava, R. K., Alpert, M. I., and Shocker, A. D. (1984). A customer-oriented approach for determining market structures. *Journal of Marketing* **48**, 32-45.
- [30] Tucker, L. R., and Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika* **28**, 333-367.

# Hierarchical clustering and the construction of (optimal) ultrametrics using $L_p$ -norms

**Lawrence Hubert**

*University of Illinois, Champaign, USA*

**Phipps Arabie**

*Rutgers University, Newark, New Jersey, USA*

**Jacqueline Meulman**

*Leiden University, Leiden, The Netherlands*

*Abstract:* The classification task of hierarchical clustering can be characterized as one of constructing for an object set  $S$  a sequence of successively less-refined partitions that attempts to represent the pattern of entries in a given symmetric proximity matrix defined between the objects. We discuss this process of constructing a partition hierarchy by the fitting through an  $L_p$ -norm (for  $p = 1, 2$ , or  $\infty$ ) of a second symmetric matrix whose entries represent what is called an ultrametric and which can be used to induce a partition hierarchy. A dynamic programming strategy, and a heuristic extension for larger object sets, is suggested as the computational mechanism for carrying out the procedure of combinatorial search for the ultrametric that is the best-fitting according to the chosen  $L_p$ -norm. A numerical example is used to illustrate the complete fitting process that relies on a proximity matrix provided. A final extension is presented for the construction of best-fitting ultrametrics based on two-mode proximity data defined between distinct object sets.

*Key words:* Ultrametric,  $L_p$ -norm, hierarchical clustering, dynamic programming, partitioning.

AMS subject classification: Primary 62H30, 92G30; secondary 90C39.

## 1 Introduction

One of the most studied data analysis topics in the field of classification is that of constructing a hierarchical clustering for an object set,  $S = \{O_1, \dots, O_n\}$ , based on some given  $n \times n$  symmetric proximity matrix  $\mathbf{P} = \{p_{ij}\}$ ; an entry  $p_{ij}$  ( $= p_{ji} \geq 0$ , and  $p_{ii} = 0$ ) is assumed to represent the dissimilarity of the objects  $O_i$  and  $O_j$ , where larger values correspond to the more dissimilar objects. A hierarchical clustering of  $S$  can be represented by a sequence of partitions,  $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_T$ , where  $\mathcal{P}_1$  is the (disjoint) partition in which each object forms its separate class,  $\mathcal{P}_T$  is the (conjoint) partition containing all objects in  $S$  within a single class, and  $\mathcal{P}_t$  is constructed by uniting two or more classes in  $\mathcal{P}_{t-1}$ . (Most commonly, only one pair of classes will be united in  $\mathcal{P}_{t-1}$ , so that  $T = n$  and  $\mathcal{P}_t$  therefore includes  $n - t + 1$  classes.) The task of hierarchical clustering is typically carried out by a greedy optimization strategy, which begins with  $\mathcal{P}_1$  and successively identifies  $\mathcal{P}_t$  from  $\mathcal{P}_{t-1}$ , for  $t \geq 2$ , by minimizing some chosen measure of proximity between the subsets that could be united to form a new class in  $\mathcal{P}_t$ . Most commercially available statistical software packages (e.g., SYSTAT, SPSS, and SAS) implement their routines for hierarchical clustering in this manner and with various choices for how the proximity between subsets might be defined.

The present paper is concerned with this particular problem of constructing a partition hierarchy that is intended to represent the patterning of relationships present in the proximity matrix  $\mathbf{P}$ , but will do so indirectly by fitting a second matrix to  $\mathbf{P} = \{p_{ij}\}$ , denoted by  $\mathbf{U} = \{u_{ij}\}$ , minimizing an  $L_p$ -norm (for one of the usual values chosen for  $p$  of 1, 2, or  $\infty$ ). The entries in the fitted matrix  $\mathbf{U}$  will satisfy a collection of linear inequality/equality constraints, characterizing what is called an ultrametric, that in turn can be used to retrieve a specific partition hierarchy for the object set  $S$ . The fitting task itself will be carried out through a recursive optimization strategy based on dynamic programming which for small object sets can provide globally optimal solutions. Later sections of the paper discuss the heuristic use of the same dynamic programming strategy for dealing with larger object sets, and an extension of the hierarchical clustering task for proximity matrices that only contain dissimilarity values between the objects from two distinct sets (i.e., two-mode proximity data).

## 2 Ultrametrics

A concept routinely encountered in formal discussions of hierarchical clustering is that of an ultrametric, which can be characterized by any non-negative  $n \times n$  symmetric dissimilarity matrix for the objects in  $S$ , denoted generically as  $\mathbf{U} = \{u_{ij}\}$ , where  $u_{ij} = 0$  if and only if  $i = j$  and the entries in  $\mathbf{U}$  satisfy the ultrametric inequality:  $u_{ij} \leq \max\{u_{ik}, u_{jk}\}$  for  $1 \leq i, j, k \leq n$ . An alternative characterization of this last inequality would be that for all distinct object triples,  $O_i$ ,  $O_j$ , and  $O_k$ , the largest two dissimilarities among  $u_{ij}$ ,  $u_{ik}$ , and  $u_{jk}$  are equal and (therefore) not smaller than the third. Any ultrametric identifies a specific partition hierarchy,  $\mathcal{P}_1, \dots, \mathcal{P}_T$ , where those object pairs defined between subsets united in  $\mathcal{P}_{t-1}$  to form  $\mathcal{P}_t$  all have a common ultrametric value that is not smaller than those for object pairs defined within these same subsets. Thus, the individual partitions in the sequence can be identified by increasing a threshold variable from zero and observing that  $\mathcal{P}_t$  is associated with a particular threshold value where all dissimilarities within a class in  $\mathcal{P}_t$  are less than or equal to this threshold and all dissimilarities between the classes in  $\mathcal{P}_t$  are strictly greater. Conversely, the collection of all ultrametric matrices can be decomposed into equivalence classes where all members of an equivalence class induce the same partition hierarchy. If  $\mathcal{P}_1, \dots, \mathcal{P}_T$  denotes the specific partition hierarchy induced by all members of an equivalence class, we will refer to one particular member of this class as the *base ultrametric* defined by  $\mathbf{U}^o = \{u_{ij}^o\}$ , where  $u_{ij}^o = \min\{t - 1 \mid \text{objects } O_i \text{ and } O_j \text{ appear within the same class in partition } \mathcal{P}_t\}$ . All members of an equivalence class can be obtained from the entries for the base ultrametric by a strictly monotonic function that maps zero to zero. Moreover, since  $\mathbf{U}^o$  contains  $T - 1$  distinct positive values, each member of this equivalence class will also contain  $T - 1$  distinct positive values, where the  $(t - 1)^{\text{st}}$  largest corresponds to partition  $\mathcal{P}_t$  in the hierarchy and is implicitly associated with those object pairs that appear together for the first time within a subset in  $\mathcal{P}_t$ .

An ultrametric matrix is a convenient device for representing in matrix form the partition hierarchy it induces, and specifically, the integer-valued base ultrametric can serve as a direct way for generating the explicit set of linear inequality/equality constraints that any ultrametric within an equivalence class must satisfy. Thus, one could find a best-fitting ultrametric within an equivalence class by fitting  $\{u_{ij}\}$  to the original proximity matrix  $\{p_{ij}\}$  through, for example, an  $L_p$ -norm regression strategy that incorporates the linear inequality/equality constraints implied by the base ultrametric (e.g., those in Späth, 1991, Chapter 5). It is also possible to



use the ultrametric notion more fundamentally as the basic mechanism for obtaining a partition hierarchy in the first place. Explicitly, we suggest here the development of hierarchical clustering methods by directly attempting to find a best-fitting ultrametric for  $\mathbf{P}$  by optimizing a loss criterion defined by an  $L_p$ -norm between  $\{p_{ij}\}$  and a (to be identified) ultrametric matrix  $\{u_{ij}\}$ . This usage of an  $L_p$ -norm is more general than what has been done thus far in the literature; the extant methods that attempt directly to obtain a best-fitting ultrametric have all adopted a least-squares criterion and some auxiliary search strategy for locating an appropriate set of constraints to impose (e.g., see Hartigan, 1967; Carroll and Pruzansky, 1980; De Soete, 1984; Chandon and De Soete, 1984; Hubert and Arabie, 1995).

To be specific, suppose for a given partition hierarchy,  $\mathcal{P}_1, \dots, \mathcal{P}_n$  (so  $T = n$ ), we let  $C_{t-1}^{(u)}$  and  $C_{t-1}^{(v)}$  denote the two classes united in  $\mathcal{P}_{t-1}$  to form  $\mathcal{P}_t$ , and specify  $b_{t-1}$  to be some appropriate aggregate (or 'average') value of the proximities for object pairs between  $C_{t-1}^{(u)}$  and  $C_{t-1}^{(v)}$ . Denoting the set of proximities between  $C_{t-1}^{(u)}$  and  $C_{t-1}^{(v)}$  as  $B_{t-1}(u, v) \equiv \{p_{i'j'} \mid O_{i'} \in C_{t-1}^{(u)}, O_{j'} \in C_{t-1}^{(v)}\}$ , and depending on the  $L_p$ -norm chosen, this between-subset aggregate value will be variously defined as the median ( $L_1$ ), the mean ( $L_2$ ), or the average of the maximum and the minimum proximities ( $L_\infty$ ) in the set  $B_{t-1}(u, v)$ . The loss functions based on an  $L_p$ -norm used to index the adequacy of a given partition hierarchy in producing an ultrametric fitted to  $\mathbf{P}$  are for the

$L_1$ -norm:

$$\sum_{t=2}^n \sum_{O_{i'} \in C_{t-1}^{(u)}, O_{j'} \in C_{t-1}^{(v)}} |p_{i'j'} - b_{t-1}|,$$

where  $b_{t-1}$  is the median proximity in the set  $B_{t-1}(u, v)$ ;

$L_2$ -norm:

$$\sum_{t=2}^n \sum_{O_{i'} \in C_{t-1}^{(u)}, O_{j'} \in C_{t-1}^{(v)}} (p_{i'j'} - b_{t-1})^2,$$

where  $b_{t-1}$  is the mean proximity in the set  $B_{t-1}(u, v)$ ;

$L_\infty$ -norm:

$$\sum_{t=2}^n \max_{O_{i'} \in C_{t-1}^{(u)}, O_{j'} \in C_{t-1}^{(v)}} |p_{i'j'} - b_{t-1}|,$$

where  $b_{t-1}$  is the average of the minimum and maximum proximities in the set  $B_{t-1}(u, v)$ .

For all three  $L_p$ -norms, an optimal ultrametric will be one for which the order constraint on the between-subset aggregate values holds:  $b_1 \leq$

$b_2 \leq \dots \leq b_{n-1}$ , and the norm is minimized. For such an optimal solution, the between-subset aggregate values,  $b_1, \dots, b_{n-1}$ , define the distinct entries in an (optimal) fitted ultrametric. (It might be noted that since less than  $n - 1$  distinct values could be identified if some of the between-subset aggregate values are tied, the search for an optimal ultrametric can assume without loss of generality that  $T = n$ , and for  $t \geq 2$ , only two classes are united within  $\mathcal{P}_{t-1}$  to form  $\mathcal{P}_t$ . Also, as a technical convenience, we allow the possibility that some of the between-subset aggregate values may be identically zero when the proximities for calculating these are all zero. Although not technically an ultrametric since zero ultrametric values should not correspond to distinct objects, its structure would still satisfy the central ultrametric inequality for distinct object triples.)

### 3 A dynamic programming strategy for identifying (optimal) ultrametrics

The optimization task of constructing optimal ultrametrics fitted to a given proximity matrix  $\mathbf{P}$  may be fairly easy to state, but the problem itself is a computationally very difficult one to solve. For both the  $L_1$ - and  $L_2$ -norm, for instance, the task has been shown to fall into the class of NP-hard problems (see Křivánek and Morávek, 1986; Křivánek, 1986; for a recent comprehensive review, see Day, 1996); thus, there is the usual expectation that for larger object sets, methods guaranteeing optimality would become computationally infeasible to implement. Keeping these computational difficulties in mind, along with the eventual necessity of moving to heuristic methods of solution for larger object sets, we will still begin with a strategy that can fit an optimal ultrametric to  $\mathbf{P}$  for each of the three  $L_p$ -norms introduced in the last section. The approach suggested is based on dynamic programming and the construction of a recursive system that will eventually produce an optimal solution. There are some complications that arise in the use of a straightforward dynamic programming formulation because of the need to impose an order constraint on the successive between-subset aggregate values, and these difficulties will be addressed below in some detail. In addition, a strategy for heuristically extending the basic dynamic programming formulation is developed in the next subsection for dealing with large(r) object set sizes.

#### 3.1 Identifying optimal ultrametrics

To implement a dynamic programming approach for locating an optimal ultrametric, we first define a collection of sets,  $\Omega_1, \dots, \Omega_n$ , where  $\Omega_k$  contains

all partitions of the  $n$  objects in  $S$  into  $n - k + 1$  classes. For convenience, a member of  $\Omega_k$  is denoted by  $A_k$ ; thus,  $\Omega_1$  contains the single partition  $A_1$  that has  $n$  classes in which each of the  $n$  objects forms a separate class, and  $\Omega_n$  contains the single partition  $A_n$  that includes one class for all of the  $n$  objects in  $S$ . We will say that a transition from  $A_{k-1} \in \Omega_{k-1}$  to  $A_k \in \Omega_k$  is *permissible* if the union of two classes in  $A_{k-1}$  produces  $A_k$ , and if an admissibility criterion to be discussed shortly is satisfied (that would [hopefully] ensure that the sequence of between-subset aggregate values is nondecreasing). A function  $\mathcal{F}(A_k)$  for  $A_k \in \Omega_k$  is defined as the optimal value for the sum of the contributions for the chosen  $L_p$ -norm up to the partition  $A_k$ . Beginning with  $\mathcal{F}(A_1) \equiv 0$  for  $A_1 \in \Omega_1$ , we construct  $\mathcal{F}(A_k)$  recursively by

$$\mathcal{F}(A_k) = \min \{ \mathcal{F}(A_{k-1}) + C(A_{k-1}, A_k) \},$$

where the minimum is taken over all  $A_{k-1} \in \Omega_{k-1}$  for which a transition is permissible to  $A_k \in \Omega_k$ , and  $C(A_{k-1}, A_k)$  is the incremental cost of transforming  $A_{k-1}$  to  $A_k$  characterized by the appropriate  $L_p$ -norm when that pair of subsets in  $A_{k-1}$  is united to form  $A_k$ . (It is this latter independence of incremental cost from how  $A_{k-1}$  was obtained that is crucial to proving the validity of the recursive process.) Finally, an optimal solution is identified by  $\mathcal{F}(A_n)$  for the single entity  $A_n \in \Omega_n$ , and a partition hierarchy attaining this optimal value identified by working backwards through the recursion starting from  $\Omega_n$  and proceeding to  $\Omega_1$  and tracing the process of how  $\mathcal{F}(A_n)$  was generated.

One unresolved issue needing discussion is the explicit imposition of some type of admissibility criterion for defining a permissible transition from  $A_{k-1}$  to  $A_k$  that could ensure a nondecreasing sequence of between-subset aggregate values. Unfortunately, the validity of the recursive process depends on the property that any proposed criterion for admissibility must only involve  $A_{k-1}$  and  $A_k$  and their relation to the matrix  $\mathbf{P}$ , and specifically *not* on how  $A_{k-1}$  may have been arrived at. Thus, it is not possible to define admissibility directly by requiring the between-subset aggregate value that defines  $A_k$  from  $A_{k-1}$  to be greater than or equal to the last between-subset aggregate value that led to  $A_{k-1}$  from  $A_{k-2}$ . What can be offered, however, are two (less-than-ideal) alternatives: (a) an admissibility criterion based only on  $A_{k-1}$  and  $A_k$  that may sometimes be too lenient and thus fail to ensure that the collection of between-subset aggregate values are nondecreasing for the (purportedly optimal) identified ultrametric, or (b) an admissibility criterion based only on  $A_{k-1}$  and  $A_k$  that may be too strict, and the (purportedly optimal) identified ultrametric could in fact not be the absolute best obtainable.

To be specific, the possibly too lenient criterion rests on the observation (made originally by Chandon, Lemaire, and Pouget, 1980, for the  $L_2$ -norm) that in an optimal ultrametric based on any of the three  $L_p$ -norms (with the notion of an aggregate value defined by the median, mean, or the average of the two extreme proximities), the nondecreasing constraint on the between-subset aggregate values,  $b_1 \leq \dots \leq b_{n-1}$ , requires that  $b_t$  be both greater than or equal to each such aggregate value calculated within a subset of  $A_{k-1}$ , and less than or equal to the aggregate value of all proximities between the subsets in  $A_k$ . Since these two conditions may be evaluated given only  $A_{k-1}$  and  $A_k$ , they can be imposed in defining whether a transition from  $A_{k-1}$  to  $A_k$  is permissible. Alternately, the possibly too strict admissibility criterion would require that  $b_t$  be less than or equal to any between-subset aggregate value calculated for the new subset formed in  $A_k$  and some other subset present in  $A_k$ . This latter criterion would ensure that no (nontrivial) order inversions in the sequence of between-subset aggregate values would exist (a trivial inversion would be one in which an inversion may be present in the collection of between-subset aggregate values, but it can be removed by a simple reordering of when two disjoint subsets are formed).

The computer program relied on for the numerical examples in Section 4 allows the imposition of either of the two admissibility criteria discussed above. As a suggested analysis strategy, one would begin with the former (and possibly too lenient) admissibility criterion and if no nontrivial order inversions in the between-subset aggregate values are found, an optimal ultrametric has been identified. If nontrivial order inversions were present, the possibly too strict admissibility criterion could be adopted, and the then identified ultrametric presumed optimal (but with the caveat that it could be possible in some [rare] instances for an even better ultrametric to be generated). (For convenience of reference, the program we use is referred to by the acronym HPHI, for ‘*H*euristic *P*rogramming *H*ierarchical clustering’, where the term ‘heuristic’ is included because of the extensions it includes for dealing with larger object sets, as discussed in the section to follow.)

### 3.2 Heuristic extensions for large(r) object sets

When the number of objects in  $S$  is even moderate in size, the random access memory storage requirements necessary for a dynamic programming approach to constructing an optimal ultrametric can become quite large. Necessary for implementing the proposed recursive strategy is the availability of large arrays associated with the sets,  $\Omega_1, \dots, \Omega_n$ , that contain for *all*

partitions of  $S$  the recursively-constructed values  $\mathcal{F}(A_k)$  for  $A_k \in \Omega_k$ , as well as a mechanism for keeping track of what previous partitions in  $\Omega_{k-1}$  led to these optimal values  $\mathcal{F}(A_k)$ .<sup>1</sup> For larger object sets, HPHI allows two options: (a) finding optimal ultrametrics for subsets of  $S$ , and (b) finding optimal ultrametrics when the basic objects to be hierarchically partitioned are themselves subsets of  $S$ . By the judicious and repeated use of these two options, we have been able to approach object sets with reasonably large sizes (and will do so for an object set of size 30 in the next section).

The analysis strategy we suggest begins by identifying [possibly through a heuristic mechanism] a partition of  $S$ , say  $\mathcal{P}_e$ , that is initially forced to be induced as part of the best-fitting ultrametric we construct. The classes of  $\mathcal{P}_e$  are first treated as the basic objects on which an ultrametric is to be obtained, i.e., we begin with the classes of  $\mathcal{P}_e$  and complete the identification of an optimal ultrametric from this point on. Secondly, each of the classes of  $\mathcal{P}_e$  is then used to obtain a separate optimal ultrametric for the objects in that class. When these results are concatenated, an optimal ultrametric is identified, subject to the constraint that  $\mathcal{P}_e$  is part of the partition hierarchy it induces. Obviously, if  $\mathcal{P}_e$  is chosen appropriately to begin with, the concatenated results would be optimal for the complete object set  $S$ . A check on the choice of  $\mathcal{P}_e$  (however it was obtained initially) can be carried out by using object classes identified within the subsets defining  $\mathcal{P}_e$  as the basic units on which an optimal ultrametric is to be constructed and then completing the fitting from this point on. If  $\mathcal{P}_e$  is retrieved as part of this latter process, some obviously increased confidence is obtained that the concatenated ultrametric may be the best we can find. If, on the other hand,  $\mathcal{P}_e$  is not retrieved, we could then repeat this same strategy with whatever partition was observed (presumably for the same number of classes as contained in  $\mathcal{P}_e$ ). This whole process could be carried out iteratively until convergence. Obviously, an absolute guarantee of optimality is not possible through this type of heuristic search, but the eventual stability achieved leads to an ultrametric that is usually very good (although not verifiably optimal). Throughout this discussion it is assumed that the subsets of objects for which separate optimal ultrametrics are generated, or the number of object classes to be used in obtaining an optimal ultrametric beginning from that point, are all of a size that could be handled optimally (i.e., some number in the lower teen's).

---

<sup>1</sup>Given the usual Pentium-level processors now commonly available and the amount of memory these systems typically contain, the program we have developed can deal (optimally) with object set sizes in the lower teen's, but even this requires the capability of Fortran90 to allocate very large arrays dynamically (and inform the user whether sufficient memory exists on the system to solve the problem of the size being requested).

## 4 A numerical illustration

To illustrate the construction of best-fitting ultrametrics based on the  $L_p$ -norm for a given proximity matrix, we use a data set originally collected by Arabie and Rips (1973) for a replication of a study initially conducted by Henley (1969) involving the subjectively-judged similarity of 30 animals. Fifty-three subjects assessed the similarity between all 435 animal pairs based on a scale from 1 (extremely dissimilar) to 10 (extremely similar). Table 1 provides the animal names and the summed ratings over the subjects subtracted from the maximum of 530 so the proximities would be keyed as dissimilarities. (We provide these data in Table 1 as a convenience to others who may wish to use this proximity matrix in their own methodological examples. Although these data have been analyzed elsewhere (see e.g., De Soete and Carroll, 1996), they have not been published explicitly.)

Based on the data of Table 1, the results are presented below for each of the three  $L_p$ -norms using the heuristic process of Section 3.1 for finding best-fitting ultrametrics. Specifically, a five-class partition,  $\mathcal{P}_e$ , of the object set  $S$  was first identified heuristically (the greedy complete-link hierarchical clustering method was used up to the level of five classes). An (optimal) ultrametric was then found for each of the five classes within  $\mathcal{P}_e$ , and based on these separate ultrametrics, a collection of (smaller) object subsets identified and treated as the starting point from which to finish the identification of an ultrametric for the complete object set  $S$ . Based on this latter ultrametric, the object classes for the induced five-class partition were then considered as defining an initial partition,  $\mathcal{P}_e$ , and the whole procedure repeated. For all three  $L_p$ -norms, the latter five-class partitions were retrieved immediately. In all of these analyses, and as suggested in the last section, the admissibility criterion that may at times be too lenient (to ensure a strictly nondecreasing between-subset collection of aggregate values) was first used, and when nontrivial order inversions were observed (as they were for a few of the analyses carried out), the more strict admissibility condition was then adopted.

The results for both the  $L_1$ - and  $L_2$ -norm are very similar, and the same five-class partition was induced for the corresponding ultrametrics:

A: {bear (2), cat (5), dog (10), fox (13), leopard (18), lion (19), tiger (28), wolf (29)} — carnivorous feline/canine animals plus the omnivorous bear

B: {beaver (3), chipmunk (7), mouse (21), rabbit (23), raccoon (24), rat (25), squirrel (27)} — small rodent-like animals

C: {antelope (1), camel (4), cow (8), deer (9), donkey (11), elephant (12), giraffe (14), goat (15), horse (17), sheep (26), zebra (30)} — large

hoofed herbivores (ungulates)

D: {chimpanzee (6), gorilla (16), monkey (20)} — primates

E: {pig (22)} — *Suidae*

The (optimal) ultrametrics defined by the values of the between-subset aggregate values for the  $L_1$ - and  $L_2$ -norm constructed within each of the classes labeled above as A, B, C, and D are given below (we also present those for the  $L_\infty$ -norm in the case of the two classes labeled B and D that were also observed in the retrieved ultrametric using this latter norm). Within each class we also provide a summary measure of the discrepancy between the proximities and fitted values by giving the contribution each class has to the overall  $L_p$ -norm measure being minimized.

| <i>level</i> | <i>new class formed</i>                                                                                           | $L_1$      | $L_2$        | $L_\infty$ |
|--------------|-------------------------------------------------------------------------------------------------------------------|------------|--------------|------------|
| A:8          | {2, 5, 10, 13, 18, 19, 28, 29}<br>an addition of omnivorous bear to the<br>carnivorous feline plus canine classes | 282.0      | 278.0        |            |
| A:7          | {5, 10, 13, 18, 19, 28, 29}<br>the union of the carnivorous feline and<br>canine classes                          | 217.0      | 222.6        |            |
| A:6          | {10, 13, 29}<br>canines                                                                                           | 104.5      | 104.5        |            |
| A:5          | {5, 18, 19, 28}<br>felines                                                                                        | 57.0       | 61.3         |            |
| A:4          | {13, 29}<br>nondomestic canines                                                                                   | 57.0       | 57.0         |            |
| A:3          | {18, 19, 28}<br>nondomestic felines                                                                               | 35.0       | 35.0         |            |
| A:2          | {18, 28}<br>feline(subclass)                                                                                      | 24.0       | 24.0         |            |
| A:1          | (all separate)<br>contribution to the norm measures:                                                              | —<br>557.0 | —<br>22,100. |            |
| B:7          | {3, 7, 21, 23, 24, 25, 27}                                                                                        | 230.5      | 212.3        | 206.5      |
| B:6          | {3, 7, 23, 24, 27}                                                                                                | 197.5      |              |            |
| B:6          | {3, 23, 24}<br>somewhat larger animals                                                                            |            | 200.5        | 200.5      |
| B:5          | {3, 7, 24, 27}                                                                                                    | 173.5      |              |            |
| B:5          | {7, 21, 25, 27}<br>very small animals                                                                             |            | 163.5        | 164.5      |
| B:4          | {3, 24}                                                                                                           | 123.0      | 123.0        | 123.0      |
| B:3          | {7, 27}                                                                                                           | 22.0       | 22.0         | 22.0       |

|      |                                          |       |          |      |
|------|------------------------------------------|-------|----------|------|
|      | long-bushy-tail animals                  |       |          |      |
| B:2  | {21, 25}                                 | 6.0   | 6.0      | 6.0  |
|      | long-naked-tail animals                  |       |          |      |
| B:1  | (all separate)                           | —     | —        | —    |
|      | contribution to the norm measures:       | 458.0 | 18,500.  | 79.5 |
| C:11 | {1, 4, 8, 9, 11, 12, 14, 15, 17, 26, 30} | 286.0 | 298.2    |      |
|      | the final addition of elephant           |       |          |      |
| C:10 | {1, 4, 8, 9, 11, 14, 15, 17, 26, 30}     | 238.0 | 242.7    |      |
| C:9  | {1, 4, 9, 11, 14, 17, 30}                | 212.5 | 215.9    |      |
| C:8  | {8, 15, 26}                              | 200.5 | 200.5    |      |
|      | farm animals                             |       |          |      |
| C:7  | {4, 14}                                  | 174.0 | 174.0    |      |
|      | African animals                          |       |          |      |
| C:6  | {1, 9, 11, 17, 30}                       | 167.5 | 174.7    |      |
|      | horse-like animals                       |       |          |      |
| C:5  | {15, 26}                                 | 93.0  | 93.0     |      |
|      | farm animals (subclass)                  |       |          |      |
| C:4  | {11, 17, 30}                             | 81.5  | 81.5     |      |
|      | equine                                   |       |          |      |
| C:3  | {1, 9}                                   | 49.0  | 49.0     |      |
|      | deer-like animals                        |       |          |      |
| C:2  | {11, 17}                                 | 31.0  | 31.0     |      |
|      | domestic equine                          |       |          |      |
| C:1  | (all separate)                           | —     | —        |      |
|      | contribution to the norm measures:       | 149.6 | 298,200. |      |
| D:3  | {6, 16, 20}                              | 49.5  | 49.5     | 49.5 |
| D:2  | {6, 20}                                  | 26.0  | 26.0     | 26.0 |
| D:1  | (all separate)                           | —     | —        |      |
|      | contribution to the norm measures:       | 19.0  | 200.0    | 9.5  |

Based on the five-classes, A, B, C, D, and E, the completions of a best-fitting ultrametric beginning from this point are given below for the  $L_1$ - and  $L_2$ -norm:

| <i>level</i> | <i>new class formed</i> | $L_1$ | $L_2$ |
|--------------|-------------------------|-------|-------|
| 5            | {A, B, C, D, E}         | 389.0 | 382.1 |
| 4            | {A, C, D}               | 375.0 |       |
| 4            | {A, C, D, E}            |       | 373.1 |
| 3            | {E, B}                  | 354.0 |       |
| 3            | {A, C, E}               |       | 357.5 |



|   |                                    |         |          |
|---|------------------------------------|---------|----------|
| 2 | {A, C}                             | 323.0   | 323.2    |
| 1 | (all separate)                     | —       | —        |
|   | contribution to the norm measures: | 10,494. | 598,600. |

For the  $L_\infty$ -norm, the five-class partition retrieved for the corresponding ultrametric differed slightly from that for the  $L_1$ - and  $L_2$ -norm and involved the placement of bear (2), elephant (12), and pig (22). Explicitly, the two classes previously labeled as B and D were again retrieved for the  $L_\infty$ -norm, but the three other classes varied slightly:

F: {antelope (1), camel (4), cow (8), deer (9), donkey (11), giraffe (14), goat (15), horse (17), pig (22), sheep (26), zebra (30)} — large hoofed herbivores including (appropriately) pig and excluding bear

G: {cat (5), dog (10), fox (13), leopard (18), lion (19), tiger (28), wolf (29)} — felines/canines only (excluding bear)

H: {bear (2), elephant (12)} — large animals

B: {beaver (3), chipmunk (7), mouse (21), rabbit (23), raccoon (24), rat (25), squirrel (27)} — small rodent-like animals

D: {chimpanzee (6), gorilla (16), monkey (20)} — primates

Using these latter five classes, the completion of a best-fitting ultrametric is given below for the  $L_\infty$ -norm; subsequently, the optimal ultrametrics within the classes labeled F, G, and H are given (those for the two classes B and D were provided previously along with the  $L_1$ - and  $L_2$ -norm results):

| <i>level</i> | <i>new class formed</i>                  | $L_\infty$ |
|--------------|------------------------------------------|------------|
| 5            | {B, D, F, G, H}                          | 343.5      |
| 4            | {D, F, G, H}                             | 335.0      |
| 3            | {F, G, H}                                | 331.0      |
| 2            | {F, H}                                   | 313.5      |
| 1            | (all separate)                           | —          |
|              | contribution to the norm measure:        | 362.0      |
| F:11         | {1, 4, 8, 9, 11, 14, 15, 17, 22, 26, 30} | 294.5      |
| F:10         | {8, 15, 22, 26}                          | 272.5      |
| F:9          | {1, 4, 9, 11, 14, 17, 30}                | 216.0      |
| F:8          | {8, 15, 26}                              | 200.5      |
| F:7          | {1, 9, 11, 17, 30}                       | 178.5      |
| F:6          | {4, 14}                                  | 174.0      |
| F:5          | {15, 26}                                 | 93.0       |
| F:4          | {11, 17, 30}                             | 81.5       |
| F:3          | {1, 9}                                   | 49.0       |
| F:2          | {11, 17}                                 | 31.0       |

|     |                                   |       |
|-----|-----------------------------------|-------|
| F:1 | (all separate)                    | –     |
|     | contribution to the norm measure: | 292.5 |
| G:7 | {5, 10, 13, 18, 19, 28, 29}       | 232.0 |
| G:6 | {10, 13, 29}                      | 104.5 |
| G:5 | {5, 18, 19, 28}                   | 63.0  |
| G:4 | {13, 29}                          | 57.0  |
| G:3 | {18, 19, 28}                      | 35.0  |
| G:2 | {18, 28}                          | 24.0  |
| G:1 | (all separate)                    | –     |
|     | contribution to the norm measure: | 91.0  |
| H:2 | {2, 12}                           | 305.0 |
| H:1 | (all separate)                    | –     |
|     | contribution to the norm measure: | 0.0   |

## 5 Constructing (optimal) ultrametrics for two-mode proximity data

The discussion of finding optimal ultrametrics has been restricted thus far to a single object set  $S$  for which a symmetric  $n \times n$  dissimilarity matrix  $\mathbf{P}$  is available. A direct extension is possible, however, to the context of a (two-mode)  $n_A \times n_B$  dissimilarity matrix  $\mathbf{Q} = \{q_{ij}\}$  defined between the objects from two distinct sets, say  $S_A = \{O_{r_1}, \dots, O_{r_{n_A}}\}$  and  $S_B = \{O_{c_1}, \dots, O_{c_{n_B}}\}$ , containing  $n_A$  and  $n_B$  objects respectively, and where  $q_{ij}$  denotes a dissimilarity between the (row) object  $O_{r_i}$  and the (column) object  $O_{c_j}$ . Specifically, a combined single object set  $S$  is first constructed as  $S \equiv S_A \cup S_B$  containing  $n \equiv n_A + n_B$  objects, and the same dynamic programming strategy for locating an (optimal) ultrametric is now applied to the single set  $S$  but with two modifications: (i) when considering the recursive process over the sets  $\Omega_1, \dots, \Omega_n$ , a transition from  $A_{k-1} \in \Omega_{k-1}$  to  $A_k \in \Omega_k$  is not permissible whenever the new subset formed in  $A_k$  would contain only objects from  $S_A$  or from  $S_B$ ; (ii) in generating the between-subset aggregate values and the contribution to the chosen norm measure for a transition from  $A_{k-1}$  to  $A_k$ , only those proximities defined between the object sets  $S_A$  and  $S_B$  are considered. Based on this strategy, the between-subset aggregate values producing the fitted values for the proximities in  $\mathbf{Q}$ , denoted generically as  $\mathbf{T} = \{t_{ij}\}$ , will satisfy the two-set ultrametric inequality (e.g., see Furnas, 1980; De Soete, DeSarbo, Furnas, and Carroll, 1984a, 1984b): for  $O_{r_i}, O_{r_{i'}} \in S_A$ , and  $O_{c_j}, O_{c_{j'}} \in S_B$ , the largest two values among  $t_{r_i c_j}, t_{r_i c_{j'}}, t_{r_{i'} c_j}$ , and  $t_{r_{i'} c_{j'}}$  are equal.

|                |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| <i>animal</i>  | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  | 15  |
| <i>name</i>    | 16  | 17  | 18  | 19  | 20  | 21  | 22  | 23  | 24  | 25  | 26  | 27  | 28  | 29  | 30  |
| antelope (1)   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| bear (2)       | 326 | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| beaver (3)     | 378 | 333 | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| camel (4)      | 225 | 334 | 410 | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| cat (5)        | 340 | 341 | 266 | 384 | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| chimpanzee (6) | 380 | 366 | 354 | 403 | 355 | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| chipmunk (7)   | 387 | 389 | 176 | 434 | 273 | 362 | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| cow (8)        | 240 | 290 | 395 | 260 | 376 | 406 | 403 | *   | *   | *   | *   | *   | *   | *   | *   |
| deer (9)       | 49  | 315 | 359 | 249 | 339 | 392 | 357 | 238 | *   | *   | *   | *   | *   | *   | *   |
| dog (10)       | 316 | 282 | 305 | 329 | 221 | 337 | 327 | 304 | 280 | *   | *   | *   | *   | *   | *   |
| donkey (11)    | 206 | 341 | 385 | 199 | 365 | 377 | 404 | 216 | 228 | 316 | *   | *   | *   | *   | *   |
| elephant (12)  | 301 | 305 | 416 | 256 | 420 | 383 | 469 | 258 | 342 | 382 | 285 | *   | *   | *   | *   |
| fox (13)       | 315 | 289 | 286 | 377 | 224 | 360 | 325 | 357 | 289 | 126 | 347 | 407 | *   | *   | *   |
| giraffe (14)   | 203 | 356 | 418 | 174 | 422 | 372 | 429 | 313 | 236 | 370 | 290 | 266 | 380 | *   | *   |
| goat (15)      | 179 | 338 | 361 | 257 | 327 | 381 | 371 | 193 | 196 | 240 | 204 | 363 | 313 | 310 | *   |
| gorilla (16)   | 379 | 251 | 400 | 378 | 411 | 40  | 412 | 397 | 407 | 369 | 382 | 347 | 362 | 347 | 412 |
| horse (17)     | 157 | 304 | 389 | 142 | 371 | 391 | 423 | 181 | 150 | 267 | 31  | 274 | 320 | 190 | 209 |
|                | 374 | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| leopard (18)   | 268 | 288 | 390 | 341 | 52  | 349 | 419 | 355 | 281 | 262 | 326 | 339 | 213 | 323 | 332 |
|                | 335 | 285 | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| lion (19)      | 276 | 257 | 381 | 327 | 75  | 365 | 435 | 322 | 296 | 263 | 323 | 307 | 213 | 331 | 325 |
|                | 323 | 247 | 38  | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| monkey (20)    | 383 | 358 | 352 | 403 | 359 | 26  | 310 | 419 | 381 | 334 | 377 | 381 | 357 | 388 | 372 |
|                | 59  | 395 | 363 | 369 | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| mouse (21)     | 434 | 436 | 261 | 439 | 336 | 388 | 164 | 416 | 386 | 349 | 412 | 473 | 372 | 453 | 386 |
|                | 472 | 440 | 430 | 457 | 385 | *   | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| pig (22)       | 410 | 356 | 350 | 395 | 359 | 408 | 394 | 269 | 368 | 299 | 347 | 371 | 362 | 410 | 284 |
|                | 404 | 344 | 400 | 389 | 401 | 375 | *   | *   | *   | *   | *   | *   | *   | *   | *   |
| rabbit (23)    | 321 | 394 | 207 | 407 | 271 | 378 | 201 | 400 | 323 | 297 | 390 | 435 | 301 | 420 | 340 |
|                | 430 | 394 | 391 | 403 | 360 | 222 | 343 | *   | *   | *   | *   | *   | *   | *   | *   |
| raccoon (24)   | 356 | 304 | 123 | 391 | 229 | 347 | 171 | 405 | 349 | 282 | 383 | 433 | 214 | 398 | 352 |
|                | 397 | 383 | 356 | 371 | 307 | 248 | 344 | 194 | *   | *   | *   | *   | *   | *   | *   |
| rat (25)       | 422 | 406 | 245 | 440 | 295 | 401 | 155 | 431 | 405 | 353 | 421 | 465 | 358 | 448 | 382 |
|                | 452 | 431 | 427 | 429 | 398 | 6   | 354 | 239 | 270 | *   | *   | *   | *   | *   | *   |
| sheep (26)     | 233 | 335 | 355 | 296 | 314 | 390 | 384 | 208 | 230 | 263 | 239 | 350 | 333 | 341 | 93  |
|                | 408 | 247 | 356 | 337 | 394 | 397 | 261 | 317 | 335 | 395 | *   | *   | *   | *   | *   |
| squirrel (27)  | 368 | 378 | 183 | 422 | 264 | 347 | 22  | 413 | 366 | 322 | 401 | 454 | 312 | 439 | 389 |
|                | 438 | 413 | 410 | 409 | 313 | 161 | 385 | 188 | 143 | 174 | 364 | *   | *   | *   | *   |
| tiger (28)     | 281 | 243 | 403 | 328 | 57  | 368 | 431 | 355 | 295 | 287 | 320 | 318 | 205 | 333 | 316 |
|                | 320 | 297 | 24  | 32  | 354 | 445 | 415 | 412 | 371 | 430 | 348 | 415 | *   | *   | *   |
| wolf (29)      | 301 | 246 | 338 | 349 | 245 | 366 | 397 | 345 | 312 | 83  | 317 | 374 | 57  | 362 | 303 |
|                | 339 | 279 | 180 | 177 | 382 | 417 | 383 | 367 | 292 | 374 | 310 | 377 | 181 | *   | *   |
| zebra (30)     | 129 | 319 | 396 | 214 | 347 | 375 | 416 | 228 | 178 | 293 | 116 | 287 | 307 | 211 | 222 |
|                | 367 | 47  | 244 | 252 | 378 | 437 | 370 | 384 | 377 | 431 | 258 | 413 | 240 | 290 | *   |

Table 1: A lower-triangular dissimilarity matrix between thirty animals based on data collected by Arabie and Rips (1973).

## References

- [1] Arabie, P., and Rips, L. (1973). A 3-way data set of similarities between Henley's animals. Unpublished manuscript, Stanford University.
- [2] Carroll, J. D., and Pruzansky, S. (1980). Discrete and hybrid scaling models. In *Similarity and Choice*, Eds. E. D. Lantermann and H. Feger, pp. 108–139. Bern: Huber.
- [3] Chandon, J.-L., and De Soete, G. (1984). Fitting a least squares ultrametric to dissimilarity data: Approximation versus optimization. In *Data Analysis and Informatics*, Ed. E. Diday, Vol.3, pp. 213–221. Amsterdam: North-Holland.
- [4] Chandon, J.-L., Lemaire, J., and Pouget, J. (1980). Construction de l'ultramétrie la plus proche d'une dissimilarité au sens des moindres carrés. *RAIRO Recherche opérationnelle* **14** 157–170.
- [5] De Soete, G., DeSarbo, W. S., Furnas, G. W., and Carroll, J. D. (1984a). The representation of nonsymmetric rectangular proximity data by ultrametric and path length tree structures. *Psychometrika* **49** 289–310.
- [6] De Soete, G., DeSarbo, W. S., Furnas, G. W., and Carroll, J. D. (1984b). Tree representations of rectangular proximity matrices. In *Trends in Mathematical Psychology*, Eds. R. Burggenhaut and V. Degreaf, pp. 377–392. New York: North-Holland.
- [7] De Soete, G. (1984). A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters* **2** 133–137.
- [8] De Soete, G., and Carroll, J. D. (1996). Tree and other network models for representing proximity data. In *Clustering and Classification*, Eds. P. Arabie, L. Hubert, and G. De Soete, pp. 157–197. Singapore: World Scientific.
- [9] Day, W. H. E. (1996). Complexity theory: an introduction for practitioners of classification. In *Clustering and Classification*, Eds. P. Arabie, L. J. Hubert, and G. De Soete, pp. 199–233. River Edge, NJ: World Scientific.
- [10] Furnas, G. W. (1980). Objects and their features: the metric representation of two class data. Unpublished doctoral dissertation, Stanford University.
- [11] Hartigan, J. A. (1967). Representation of similarity matrices by trees. *J. Am. Statist. Assoc.* **62** 1140–1158.
- [12] Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior* **8** 176–184.
- [13] Hubert, L. J., and Arabie, P. (1995). Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *British*

- Journal of Mathematical and Statistical Psychology* **48** 281–317.
- [14] Křivánek, M. (1986). On the computational complexity of clustering. In *Analyse des Données et Informatique*, Eds. E. Diday et al., Vol. 4, pp. 89–96. Amsterdam: North-Holland.
- [15] Křivánek, M., and Morávek, J. (1986). NP-hard problems in hierarchical clustering. *Acta Informatica* **23** 311–323.
- [16] Späth, H. (1991). *Mathematical Algorithms for Linear Regression*. New York: Academic Press.

# **$L_1$ and $L_2$ approximation clustering for mixed data: Scatter decompositions and algorithms**

**Boris Mirkin**

*Rutgers University, Piscataway, USA and*

*Central Economics-Mathematics Institute, Moscow, Russia*

*Abstract:* Clustering is considered usually an art rather than a science because of lacking comprehensive mathematical theories in the discipline. The major issue raised in this paper is that  $L_2$  and  $L_1$  approximation bilinear clustering can provide a theoretical framework for an extensive part of partitioning and hierarchic clustering concerning its algorithmical and interpretational aspects, which is supported with a theoretical evidence.

*Key words:* Partitioning, hierarchy, mixed data, approximation, contingency coefficients.

AMS subject classification: 62H30, 90C27, 05C50, 05C05.

## **1 Introduction**

Clustering is considered usually an art rather than a science because of lacking comprehensive mathematical theories in the discipline. The major issue raised in this paper is that approximation bilinear clustering can provide a theoretical framework for a part of partitioning and hierarchic clustering concerning its algorithmical and interpretational aspects. Two approximation norms,  $L_1$  and  $L_2$ , are considered and compared.

The remainder consists of two parts devoted respectively to partitioning (Sections 2 and 3) and hierarchic clustering (Section 4), and a conclusion (Section 5). In Section 2, a bilinear model relating data to a partition is considered. The model is introduced in Section 2.1 where two model-based principles for data standardization are suggested. In Section 2.2., an  $L_2$  decomposition of the data scatter into explained and unexplained parts is

discussed, especially in its relation to the nominal data case. It appears, some most known contingency measures, as Pearson chi-square, can be interpreted as contributions to the data scatter. Similar work is done for  $L_1$  in Section 2.3.

In Section 3, clustering algorithms are discussed for both,  $L_1$  and  $L_2$ , criteria. In Section 3.1, K-Means and principal cluster analysis methods are considered as locally optimal approximation techniques (the latter can be applied also for finding overlapping clusters). In Section 3.2, a prerequisite for this is outlined: interrelation between six otherwise independent parameters of cluster structure, emerging in the context of the bilinear model.

In Section 4, hierarchic clustering is put in the bilinear modeling framework. In Section 4.1, 3-valued nest indicator functions are introduced to provide for exact embedding of binary hierarchies into linear subspaces. The case of  $L_2$  hierarchic clustering is considered in Section 4.2, which is proved similar to the case of  $L_2$  partitioning except for that here decomposition concerns not only the data scatter but also the data entries and between-variable correlations. The case of  $L_1$  hierarchic clustering is treated in Section 4.3. Due to the fact that the split cluster centers must be interrelated here, the alternating minimization technique produces a modified clustering approach.

## 2 Bilinear partition model and scatter decomposition

### 2.1 Bilinear model and standardization of mixed data

Let us consider an entity-to-variable data table in which a quantitative variable  $k$  is represented by a quantitative  $N$ -dimensional column-vector  $x_k$  of its values on  $N$  entities under consideration. A binary variable (category)  $k$  is basically a question admitting only answers Yes or No for each of the entities; the values are coded 1 (Yes) and 0 (No), which produces a zero-one  $N$ -dimensional column vector  $x_k$ . A nominal variable  $k$  is coded by a zero-one  $N \times \#k$  submatrix  $x_k = (x_{iv})$  where  $\#k$  is the number of categories  $v \in k$  and  $x_{iv}$  equals 1 when entity  $i$  belongs to category  $v$  of  $k$ , and 0 otherwise.

Encoded this way, the data matrix will be denoted by  $X = (x_{iv})$  where  $i \in I$  are entities and  $v \in V$  are variables/categories corresponding to columns. These data are preprocessed into matrix  $Y = (y_{iv})$  by the stan-

dard preliminary transformation (standardization) so that

$$y_{iv} = \frac{x_{iv} - a_v}{b_v}, \quad i \in I, \quad v \in V \quad (1)$$

assuming change of both the scale factor (dividing by  $b$ ) and the origin (adding of  $a$ ) in the original column  $x_v$ . Choice of  $a$  and  $b$  as well as their meaning for categories will be discussed below after introducing the bilinear clustering model.

Let the entities be assigned into groups (clusters) presented by an additive type cluster structure which is a set of  $m$  clusters, any cluster  $t$ ,  $t = 1, \dots, m$ , being defined with two objects: 1) its membership function  $z_t = (z_{it}), i \in I$ , where  $z_{it}$  is 0 or 1 characterizing thus a cluster set  $S_t = \{i \in I : z_{it} = 1\}$ , 2) its standard point, or centroid vector,  $c_t = (c_{tv}), v \in V$ , to be combined in an  $N \times |V|$  cluster-type matrix with elements  $\sum_{t=1}^m c_{tv} z_{it}$ . ( $|V|$  is the number of columns in  $X$ .)

The cluster-type matrix models the given matrix  $Y$  via equations

$$y_{iv} = \sum_{t=1}^m c_{tv} z_{it} + e_{iv} \quad (2)$$

where residual values  $e_{iv}$  show difference between the data and the clusters. When clusters are not given a priori, they can be found in such a way that the residuals are made as small as possible, thus minimizing  $\Phi(\{|e_{iv}|\})$  where  $\Phi$  is an increasing monotone function of its arguments. The equations in (2) along with criterion  $\Phi$  to be minimized by unknown parameters,  $c_{tv}, z_{it}, e_{iv}$ , for  $y_{iv}$  given, will be referred to as the *bilinear clustering model*. This model was suggested by the author as an extension of a version of the principal component analysis technique in Mirkin (1987) and updated in Mirkin (1990). It was considered also in Chaturvedi and Carroll (1994). A detailed account of the model and its use in hard and fuzzy clustering and machine learning can be found in Mirkin (1996).

Though the model is quite similar to that of the principal component analysis (the only difference is that the “components”  $z_t$  are Boolean, not arbitrary, vectors), it has a meaning on its own, just as a clustering model. When the clusters are required to be nonoverlapping, the type-cluster matrix  $\sum_{t=1}^m c_{tv} z_{it}$  has especially simple structure considered also by Van Buuren and Heiser (1989): its rows are the vectors  $c_t = (c_{tv})$  so that every  $i$ -th row equals  $c_t$  for that specific cluster  $t$  which contains the entity  $i \in I$ .

Two Minkowski forms of criterion  $\Phi$  for minimizing the residuals are  $L_2 = \sum_{i \in I} \sum_{v \in V} e_{iv}^2$  and  $L_1 = \sum_{i \in I} \sum_{v \in V} |e_{iv}|$ . With the non-overlapping



restriction, the criteria become especially simple:

$$L_p = \sum_{i \in I} \sum_{v \in V} |y_{iv} - \sum_t c_{tv} z_{it}|^p = \sum_{v \in V} \sum_{t=1}^m \sum_{i \in S_t} |y_{iv} - c_{tv}|^p \quad (3)$$

which shows that  $L_p$ , actually, is  $L_p = \sum_{t=1}^m \sum_{i \in S_t} d^p(y_i, c_t)$  where  $d^p$  is the  $p$ -th power of the Minkowski distance.

Due to formula (3), when the membership functions are given, the optimal  $c_{tv}$  is determined only by the values  $y_{iv}$  within  $S_t$ . In particular, the least-squares ( $p = 2$ ) optimal  $c_{tv}$  is the average of  $y_{iv}$  in  $S_t$ ,  $c_{tv} = \sum_{i \in S_t} y_{iv} / |S_t|$ , while the least-moduli ( $p=1$ ) optimal  $c_{tv}$  is a median of  $y_{iv}, i \in S_t$ .

The criterion  $\Phi$ , when its argument is the data matrix  $Y = (y_{iv})$  itself,  $\Phi(\{|y_{iv}|\})$ , may be considered as a measure of the scatter of the data while  $\Phi(\{|e_{iv}|\})$  as a measure of the “unexplained” scatter. Indeed, their difference,  $\bar{\Phi} = \Phi(\{|y_{iv}|\}) - \Phi(\{|e_{iv}|\})$  will be nonnegative for any appropriate minimizer of  $\Phi$  since  $e_{iv} = y_{iv}$  (for all  $i, v$ ) and thus  $\bar{\Phi} = 0$  when all  $c_{tv} = 0$  which is not an optimal solution. Value  $\bar{\Phi}$  can be interpreted as the “explained” part of the data scatter  $\Phi(\{|y_{iv}|\})$ , which gives a decomposition of the data scatter in the two parts,  $\Phi(\{|y_{iv}|\}) = \bar{\Phi} + \Phi(\{|e_{iv}|\})$ .

In this setting, it is the data scatter which is decomposed into explained and unexplained parts due to the bilinear model; moreover, the unexplained part is nothing but the minimized criterion of the model. This is why the present author considers the data scatter as the base for choosing the data standardization parameters in (1).

Let us require that all the variables are standardized so that their contributions to the data scatter are equal to each other. The principle should be considered as an adequate formalization of the requirement of equal weight of the variables in numerical taxonomy (Sneath and Sokal, 1973). The choice of parameter  $a_v$  does not affect the model (2) for a non-overlapping cluster structure, however when the bilinear model is set forth in a sequential way with the “component” axes  $z_t$  identified one-by-one, not simultaneously, the solution heavily depends on the origin of the variable/category space. To adjust to this kind of principal/correspondence analysis methods, let us postulate an analogue to the law of minimum moment of inertia in mechanics: the origin of the variable space should be a minimizer of the data scatter.

The two scatter-based principles make the parameters defined unambiguously for  $L_1$  and  $L_2$ . When  $p = 2$ , they lead to the usual  $z$ -score standardization rule: the origin is the grand mean while the standard deviation is the scale factor, which will be referred to as *square-scatter standardization*. When  $p = 1$ , the origin must be grand median while the scale

factor is the absolute deviation, which will be referred to as *module-scatter standardization*.

In the case of mixed data, the average of a category  $v \in V$  column vector is equal, obviously, to the relative frequency of the category in  $I$ ,  $p_v$ , while its median may be 1,  $1/2$ , or 0 depending on whether  $p_v$  is larger than, equal to, or smaller than  $1/2$ , respectively. To satisfy the principle of equal contribution with  $a_v = p_v$ , the  $L_2$ -based scale factor of a category  $v$  can be taken as  $b_v = \sqrt{1 - \sum_v p_v^2}$  where summation is made by all the categories of a variable  $k$ ,  $v \in k$  (the square root of Gini index). There can be also other standardizing options suggested as, for instance,  $b_v = \sqrt{(\#k - 1)p_v}$  which is category-specific.

The absolute deviation of the values of a binary column vector from the median is equal to  $p_v$  or  $1 - p_v$  depending on whether  $p_v$  is less than  $1/2$  or not.

## 2.2 Decomposition of the least-squares criterion

With the least-squares criterion, the following decomposition holds (see, for instance, Jain and Dubes, 1988).

**Statement 1** *If values  $c_{tv}$  are optimal for a partition  $S = \{S_t\}$  of  $I$ , then*

$$\sum_{i \in I} \sum_{v \in V} y_{iv}^2 = \sum_{t=1}^m \sum_{v \in V} c_{tv}^2 |S_t| + \sum_{i \in I} \sum_{v \in V} e_{iv}^2, \quad (4)$$

Usually the equation in (4) is interpreted in terms of analysis of variance. In cluster analysis, interpretation of (4) in terms of the contributions to data scatter seems more helpful. The contribution of a pair variable-cluster  $(v, t)$  to the explained part of the data scatter is  $c_{tv}^2 |S_t|$ : it is proportional to the cluster cardinality and to the squared distance from the grand mean of the variable to its mean (standard value) within the cluster. The contribution of an entity-cluster pair can be evaluated as  $(y_i, c_t)$  because  $c_{tv}^2 |S_t| = (\sum_{i \in S_t} y_{iv} / |S_t|) c_{tv} |S_t| = \sum_{i \in S_t} y_{iv} c_{tv}$ . These cluster-specific salience weights of the variables and entities can be employed for concept learning and feature selection in machine learning (Mirkin, 1997).

To analyze the contributions of nominal variables and their categories to the scatter part explained via cluster partition  $S$ , let us denote the relative frequency (proportion of ones) of category  $v$  in set  $I$  by  $p_v$  and the proportion of entities simultaneously having category  $v$  and belonging to cluster  $S_t$ , by  $p_{vt}$ . Then, for any category  $v$  standardized by formula (1), its mean within cluster  $S_t$  is equal to  $c_{tv} = (p_{vt} - p_t a_v) / (p_t b_v)$ . The contribution of a category-cluster pair  $(v, t)$  to the explained part of the

data scatter is equal to

$$s(v, t) = c_{tv}^2 |S_t| = N(p_{vt} - p_t a_v)^2 / (p_t b_v^2), \quad (5)$$

which can be considered a measure of association between category  $v$  and cluster  $t$ .

Since every nominal variable  $k$  is considered as the set of its categories  $v$ , the joint contribution of  $k$  and the set of the clusters  $S_t$  to the scatter of the data is equal to  $F(k, S) = \sum_t \sum_{v \in k} s(v, t)$  which is

$$F(k, S) = N \sum_{t=1}^m \sum_{v \in k} \frac{(p_{vt} - p_t a_v)^2}{p_t b_v^2} \quad (6)$$

by (5). Substituting the appropriate values of  $a_v = p_v$  and  $b_v$ , we arrive at the following.

**Statement 2** *For criterion  $L_2$ , the contribution of a nominal variable  $k \in K$  to the part of the square scatter of the square standardized data that is explained by the (sought or found or expert-given) cluster partition  $S = \{S_1, \dots, S_m\}$ , is equal to*

$$\Delta(S/k) = N \sum_{v \in k} \sum_{t=1}^m \frac{(p_{vt} - p_v p_t)^2}{p_t} \quad (7)$$

when  $b_v = 1$  (no normalization), or

$$W(R/k) = N \sum_{v \in k} \sum_{t=1}^m \frac{(p_{vt} - p_v p_t)^2 / p_t}{1 - \sum_{v \in k} p_v^2} \quad (8)$$

when  $b_v = \sqrt{1 - \sum_{v \in k} p_v^2}$  (a standardizing option suggested), or

$$M(S/k) = \frac{N}{\#k - 1} \sum_{v \in k} \sum_{t=1}^m \frac{(p_{vt} - p_v p_t)^2}{p_v p_t} \quad (9)$$

when  $b_v = \sqrt{p_v(\#k - 1)}$  (another standardizing option).

All three of the coefficients relate to well known indices of contingency between the nominal variables:  $M(S/k)$  is a normalized version of the Pearson chi-square coefficient,  $\Delta(R/k)$  is proportional to the coefficient of reduction of the error of proportional prediction, and  $W(R/k)$  is the Wallis coefficient. Amazingly, it is the method of data standardization which determines which of the coefficients is produced as the contribution-to-scatter.

The contribution of a quantitative variable into the explained part of the  $L_2$  data scatter is also meaningful. When the variable  $k$  is standardized, it is exactly  $N\eta^2(k, S)$  where  $\eta^2(k, S)$  is the so-called correlation ratio (squared).

### 2.3 Least-moduli decomposition

Similar decomposition can be done for the least-moduli criterion (the contents of this section is a corrected version of section 6.1.4 in Mirkin, 1996).

Let  $S_t$  be an entity subset, and  $S_{tv} = \{i \in S_t : |y_{iv}| < |c_{tv}| \text{ \& } \text{sgn } y_{iv} = \text{sgn } c_{tv}\}$  where, as usual,  $\text{sgn } x$  is 1 if  $x > 0$ , 0 if  $x = 0$ , and  $-1$  if  $x < 0$ . This means that  $S_{tv} = \{i \in S_t : 0 \leq y_{iv} < c_{tv}\}$  if  $c_{tv}$  is positive or  $S_{tv} = \{i \in S_t : c_{tv} < y_{iv} \leq 0\}$  if  $c_{tv}$  is negative. Having a value  $c_{tv}$  fixed, the set  $S_t$  is partitioned into three subsets by the variable/category  $v$  depending on relations between  $y_{iv}$ ,  $i \in S_t$ , and  $c_{tv}$ . For  $c_{tv} > 0$ , let us denote the cardinalities of the subsets where  $y_{iv}$  is larger than, equal to or less than  $c_{tv}$  by  $n_{tv1}$ ,  $n_{tv2}$  and  $n_{tv3}$ , respectively. Then, let  $n_{tv} = n_{tv1} + n_{tv2} - n_{tv3}$ . For  $c_{tv} < 0$ , the symbols  $n_{tv1}$  and  $n_{tv3}$  are interchanged. If  $c_{tv}$  is the median of values  $y_v$  in  $S_t$  and all the values  $y_{iv}$ ,  $i \in S_t$ , are different, then  $n_{tv1} = n_{tv3}$  and  $n_{tv} = n_{tv2} = 0$  or  $= 1$  depending on the cardinality of  $S_t$  (even or odd, respectively).

**Statement 3** *When values  $c_{tv}$  are  $L_1$ -optimal for a partition  $S = \{S_t\}$  of  $I$ , the following decomposition of the module data scatter holds:*

$$\sum_{i \in I} \sum_{v \in V} |y_{iv}| = \sum_{v \in V} \sum_{t=1}^m (2 \sum_{i \in S_{tv}} |y_{iv}| + n_{tv} |c_{tv}|) + \sum_{i \in I} \sum_{v \in V} |e_{iv}|. \quad (10)$$

The proof is based on the following equation,  $|a - b| = |a| + |b| - |\text{sgn } a + \text{sgn } b| \min(|a|, |b|)$ , which holds for any real  $a$  and  $b$ .

Let us denote the contribution of a variable-cluster pair  $(v, t)$  to the module scatter in (10) by  $s(t, v) = 2 \sum_{i \in S_{tv}} |y_{iv}| + n_{tv} |c_{tv}|$ . Based on this, various relative contribution measures can be defined: (a) variable to scatter,  $w(v) = \sum_t s(t, v) / \sum_{i,v} |y_{iv}|$ ; (b) cluster to scatter,  $w(t) = \sum_v s(t, v) / \sum_{i,v} |y_{iv}|$ ; (c) variable to cluster,  $w(v/t) = s(t, v) / \sum_v s(t, v)$ ; (d) entity to cluster,  $w(i/t) = |\text{sgn } y_{iv} + \text{sgn } c_{tv}| \min(|y_{iv}|, |c_{tv}|) - |c_{tv}|$ .

Let us consider the case when  $v$  is a category.

**Statement 4** *For any category  $v$  standardized (with arbitrary  $a_v$  and  $b_v$ ), its median,  $c_{tv}$ , in cluster  $S_t$  is equal to  $-a_v/b_v$ ,  $(1-2a_v)/2b_v$ , or  $(1-a_v)/b_v$  depending on  $p_{tv}$  is smaller than, equal to, or greater than  $0.5p_t$ , respectively. The contribution of a category-cluster pair,  $(v, S_t)$ , to the module data scatter is equal to*

$$s(t, v) = N |2p_{vt} - p_t| |c_{tv}|.$$

**Proof:** The formulas for  $c_{tv}$  are evident. To derive the formula for  $s(t, v)$ , let us see that  $S_{tv} = \emptyset$  since the values  $c_{tv}$  and  $y_{iv}$  must have different signs

if they are not equal to each other ( $y_{iv}$  may have one of two values only since  $v$  is a category). Then,  $n_{tv1} + n_{tv2} = Np_{tv}$  and  $n_{tv3} = N(p_t - p_{tv})$  when  $c_{tv} = (1 - a_v)/b_v > 0$  where  $a_v, b_v$  are the values used in the module-scatter standardization rule. Analogously,  $n_{tv1} + n_{tv2} = N(p_t - p_{tv})$  and  $n_{tv3} = Np_{tv}$  when  $c_{tv} = -a_v/b_v < 0$ .  $\square$

Putting the module-scatter based  $a_v$  and  $b_v$  into  $s(t, v)$ , we have:

**Statement 5** *The contribution of a nominal variable  $k$  to the absolute scatter of the module-scatter standardized data, as explained by the partition  $S = \{S_1, \dots, S_m\}$ , is equal to*

$$A(S/k) = \frac{N}{\#k} \left( \sum_{(v,t) \in A_+} \frac{2p_{vt} - p_t}{p_v} + \sum_{(v,t) \in A_-} \frac{p_t - 2p_{vt}}{1 - p_v} + \sum_{(v,t) \in A_=} |2p_{vt} - p_t| \right) \quad (11)$$

where  $A_+ = \{(v, t) : p_v < 0.5 \text{ and } p_{vt}/p_t > 0.5\}$ ,  $A_- = \{(v, t) : p_v > 0.5 \text{ and } p_{vt}/p_t < 0.5\}$ , and  $A_= = \{(v, t) : p_v = 0.5\}$ .

The coefficient  $A(S/k)$  takes into account the situations when the patterns of occurrences of the categories  $v \in k$  in the clusters  $t$  differ from those in the entire set  $I$ . Such a difference appears when  $v$  is frequent in  $S_t$  ( $p(v/t) > 0.5$ ) and rare in  $I$  ( $p_v < 0.5$ ), or, conversely,  $v$  is rare in  $S_t$  and frequent in  $I$ .

### 3 K-Means and bilinear partitioning

#### 3.1 Principal clustering and K-Means

Following the standard strategy of sequential extraction of factors in principal component analysis, the clusters can be extracted one by one in (2), which constitutes the method of principal cluster analysis (PCL) (applicable in both overlapping and non-overlapping cluster cases).

1. Set  $t = 1$  and define data matrix  $Y_t$  as the initial data matrix standardized according to the criterion,  $L_1$  or  $L_2$ , chosen as described in Section 2.1. Choose whether the clusters to be found are required to be nonoverlapping or they may overlap each other.

2. For  $Y = Y_t$  find a principal cluster minimizing  $L_1$  or  $L_2$  as described in the algorithm for Single Cluster Clustering (SCC) below. Define  $z_t, c_t$  as the cluster solution found (membership function and the standard point, respectively); compute its contribution to the data scatter.

3. Stop-Condition. If there must be nonoverlapping hard clusters, check whether there are yet unclustered entities remaining. (In the other case,

check the standard contribution-based stopping rule of the principal component analysis). If yes, go to 4; else end.

4. Compute the residual data  $y_{iv}^{(t+1)} = y_{iv}^t - c_{tv}z_{it}$ . In the nonoverlapping case, set  $Y_{t+1}$  by removing from  $Y_t$  all the rows corresponding to the previously found cluster  $t$ . Increase  $t$  by 1, and go to 2.

The PCL algorithm can be rephrased in terms of K-Means method (MacQueen (1967), Jain and Dubes (1988)), which, in its “parallel” version, starts with  $m$  somehow selected tentative standard points or “seeds”,  $c_t$ . Then the algorithm repeatedly performs the following two-step iteration : (1) update the partition based on the standard points : given  $c_t$ , make each  $S_t$  the set of  $y_i$  that are nearest to  $c_t$ ,  $t = 1, \dots, m$ ; (2) update the standard points: when all  $S_t$  are given, compute  $c_t$  as the mean (or median, for  $L_1$ ) of the within-cluster vectors. This algorithm is, in fact, a version of the alternating minimization for criteria  $L_p$  : (1) given  $c$ , find optimal  $z$ ; (2) given  $z$ , find optimal  $c$ .

The principal cluster analysis can be considered as a technique that exploits many of the same mechanisms, but which mitigates the need for prior knowledge, and separates clusters from the set of instances one by one. First, an initial cluster  $S_1 \subset I$  is extracted with its standard point  $c_1$ ; the complementary set represents the main “body” of entities, which serves as the source for separating additional clusters one by one. This is reflected in that fact that the main body’s standard point is fixed at 0, given the square or module scatter standardization, and it is not changed during the entire clustering computation. The algorithm SCC (Single Cluster Clustering) for separating a principal cluster at the Step 2 of PCL is as follows (the data matrix is denoted by  $Y$ , not  $Y_t$ ):

Step 1. (Selection of an extreme point). Pick a point,  $y_{i^*}$ , maximizing distance  $d(0, y_i)$ ,  $i \in I$ , from the origin (the distance is taken according to the criterion,  $L_1$  or  $L_2$ , selected : city-block or Euclidean squared). Take  $c = y_{i^*}$  as the initial center (seed) of the cluster to be found.

Step 2. (Updating of the cluster). Define cluster  $S$  of points  $y_i$  around the center  $c$  as  $S = \{i : d(y_i, c) < d(y_i, 0)\}$ .

Step 3. (Updating of the standard point). Compute the center of  $S$ ,  $c = c(S)$ , which is the median or average vector, depending on  $L_1$  or  $L_2$  is utilized.

Step 4. (Stop condition) Compare  $S$  with that at the previous iteration. If there is no difference, the process ends:  $S$  and  $c(S)$  are the result. Else go to Step 2.

A general property is that the size of an SCC-designed cluster depends

on its distance from the origin (which is just a reference point): the nearer to that point, the less the diameter of the cluster! Thus, SCC could be modified to allow the user to specify the reference-point origin based on the user’s knowledge of the variable space: the better the knowledge, the smaller the classes.

The principal cluster analysis method can be used as an option in extending the K-Means method for a wider class of situations when the user can fix a few or none tentative centers even if she/he does not know the total number of the clusters or the total number is larger than the number of tentative centers the user is able to specify (Mirkin (1996)).

**3.2 How K-Means parameters should be chosen**

The user of the K-Means method faces, usually, problems in choosing the following five important kinds of parameter associated with the method: 1) preliminary transformation of the raw data  $X$  into matrix  $Y$  to be processed; 2) entity-to-center distance  $d(x, c)$ ; 3) centroid concept; 4) number of clusters; 5) initial centers. Traditionally, these parameters are considered

| Criterion     | Data Scatter   | Metric     | Centroid | Scale Parameter    | Shift Parameter |
|---------------|----------------|------------|----------|--------------------|-----------------|
| Least Squares | Square         | Euclidean  | Average  | Standard Deviation | Average         |
| Least Moduli  | Absolute Value | City-Block | Median   | Absolute Deviation | Median          |
| Least Maximum | Maximal Range  | Chebyshev  | Midrange | Half-range         | Midrange        |

Table 1: Correspondence between clustering parameters due to the bilinear model.

as completely independent except for the obvious equality of the numbers of clusters and centers. The bilinear clustering model suggests that there is no independence anymore: the parameters are associated to the criterion for model fitting. The correspondence is presented in Table 1; The Chebyshev least-maximum criterion,  $L_\infty = \max_{i,v} |y_{iv} - \sum_t c_{tv} z_{it}|$ , also is included since all these parameters can be derived from it, too.

Table 1 can be used for determining all of the six parameters when the user is able to choose at least one of them. If, for instance, the user prefers medians as the centroids, she/he is restricted, due to the bilinear model, with the least-moduli criterion along with the city-block distance, etc.

## 4 Bilinear hierarchic clustering

### 4.1 Linear embedding binary hierarchies

To discuss hierarchic clustering, we consider a binary hierarchy as a set of subsets  $S_W = \{S_w : S_w \subseteq I, w \in W\}$  called clusters containing all singletons and  $I$  so that the clusters  $S_w$ ,  $w \in W$ , are nested and every non-singleton cluster  $S_w$ ,  $w \in W$ , is a union of its two children clusters  $S_{w1}, S_{w2} \in S_W$ .

For any nonsingleton cluster  $S_w = S_{w1} \cup S_{w2}$  ( $w, w1, w2 \in W$ ) of  $S_W$ , its three-valued *nest indicator function*  $\phi_w = (\phi_{iw})$  is defined by  $\phi_{iw} = a_w$  if  $i \in S_{w1}$ ,  $= -b_w$  if  $i \in S_{w2}$ , and  $= 0$  if  $i \notin S_w$ , where the values  $a_w$  and  $b_w$  satisfy the two conditions: (1) vector  $\phi_w$  is centered; (2) vector's  $\phi_w$  norm is 1. It is easy to see that

$$a_w = \sqrt{\frac{n_{w2}}{n_{w1}n_w}}, \text{ and } b_w = \sqrt{\frac{n_{w1}}{n_{w2}n_w}} \quad (12)$$

where  $n_w$ ,  $n_{w1}$ , and  $n_{w2}$  are cardinalities of  $S_w$  and its two children,  $S_{w1}$  and  $S_{w2}$ , respectively.

It turns out, vectors  $\phi_w$  are mutually orthogonal,  $(\phi_w, \phi_{w'}) = 0$ , which is trivial when  $S_w \cap S_{w'} = \emptyset$  and also true when  $S_w \cap S_{w'} \neq \emptyset$  since in the latter case one of the clusters is a part of the other and, thus, its components are non-zero when the other vector's components are constant. Therefore, the set  $\{\phi_w : w \in W\}$  is an ortho-normal basis of the  $(N - 1)$ -dimensional space of all  $N$ -dimensional centered vectors, and any column-centered data matrix  $Y$  can be decomposed as follows:

$$Y = \Phi C \quad (13)$$

where  $\Phi = (\phi_{iw})$  is the  $N \times (N - 1)$  matrix of the values of the nest indicator functions and  $C = (c_{wv})$  is an  $(N - 1) \times |V|$  matrix.

Since  $\Phi^T \Phi$  is the identity matrix, multiplying equality in (13) by  $\Phi^T$  leads to  $C = \Phi^T Y$ , that is,

$$c_{wk} = \sqrt{\frac{n_{w1}n_{w2}}{n_w}}(y_{w1v} - y_{w2v}) = \sqrt{\frac{n_{w1}n_w}{n_{w2}}}(y_{w1v} - y_{wv}), \quad (14)$$

where  $y_{wv}$ ,  $y_{w1v}$  and  $y_{w2v}$  are the averages of the variable/category  $v \in V$  in  $S_w$ ,  $S_{w1}$  and  $S_{w2}$ , respectively. By analogy with the factor loadings in the principal component analysis, the entries of  $C$  can be referred to as the cluster loadings.

Let us denote by  $y_w$  the  $m$ -dimensional vector of the averages of the variables in a subset  $S_w$ ,  $w \in W$ . The equality in (14) implies that both



$L_1$  and  $L_2$  norms of vector  $c_w = (c_{wk})$  can be expressed as

$$\mu_w = \sqrt{\frac{n_{w1}n_{w2}}{n_w}} d(y_{w1}, y_{w2}) \quad (15)$$

where  $d(x, y)$  is the city-block/Euclidean distance between vectors  $x, y$ . The value  $\mu_w$  is positive if  $x \neq y$ , and zero if  $x = y$ . It is an analogue of the singular value in decomposition (13) considered as an analogue of the singular-value decomposition.

Another useful property of decomposition (13) is that  $Y^T Y = C^T C$ , which is a decomposition of the between-variable covariance (or correlation) coefficients by clusters of the hierarchy  $S_W$ .

Thus, in the case when all the cluster hierarchy is available, there is not much difference between  $L_1$  and  $L_2$  cases, just the choice of distance must be done accordingly while the average serves as the center in both cases.

## 4.2 Least-squares hierarchy fitting

When the hierarchy is partly unknown so that only upper clusters are given, the exact equality in (13) must be changed for the bilinear model equation, in this case  $Y = \Phi C + E$ , where the residuals  $E$  are to be minimized. It is not difficult to prove that, when some columns of  $\Phi$  are given (as a part of a basis), the least-squares estimator for corresponding  $C$  still satisfies equation (4.1). Moreover, the data scatter is decomposed as follows:

$$\sum_{i \in I, v \in V} y_{iv}^2 = \sum_{t=1}^m \mu_t^2 + \sum_{i \in I, v \in V} e_{iv}^2 \quad (16)$$

so that finding an optimal  $m$ -column  $\Phi$  requires maximizing  $\sum_{t=1}^m \mu_t^2$ .

In the framework of the principal component analysis-like sequential fitting strategy, splitting are to be done sequentially, starting with the all set  $I$ , each time maximizing corresponding  $\mu_w^2$ . It is exactly the criterion

$$\mu_w^2 = \frac{n_{w1}n_{w2}}{n_w} d^2(y_{w1}, y_{w2}), \quad (17)$$

suggested by Edwards and Cavalli-Sforza (1965) for divisive clustering, to be maximized by splitting a cluster  $S_w$  into  $S_{w1}$  and  $S_{w2}$ . The step of taking residual data in the principal component analysis-like strategy can be skipped here since it doesn't affect the results, as is not difficult to prove. The standard K-Means method (with two clusters) can be applied as an alternating maximization technique since criterion (17) is equivalent to the least-squares clustering criterion.

The other expression in (14) leads to the same criterion expressed as  $\mu_w^2 = N_w N_{w1} d^2(y_{w1}, y_w) / N_{w2}$  which implies a different, SCC-like algorithm, because the center  $y_w$  of  $S_w$  does not vary in the splitting process.

### 4.3 Least-moduli hierarchic clustering

The least moduli estimator of  $C$  in the bilinear equation  $Y = \Phi C + E$  (with  $E$  minimized) does not satisfy (14) anymore. Let us discuss the sequential fitting strategy with the criterion,  $L_1 = \sum_{i \in I} \sum_{v \in V} |y_{iv} - c_v \phi_i|$ , to be minimized by unknown  $c_v$ ,  $\phi_i$  (index  $w$  is omitted, for notational simplicity). Let us denote  $c = (c_v)$  and apply definition of  $\phi$ ; the criterion becomes:

$$L_1(S_1, S_2, c) = \sum_{i \in S_1} d(y_i, c') + \sum_{i \in S_2} d(y_i, c'') \quad (18)$$

where  $d$  is the city-block distance and  $c' = (n_2/nn_1)^{1/2}c$ ,  $c'' = -(n_1/nn_2)^{1/2}c$ . The latter equations give  $n_1c' = -n_2c''$ , which makes the alternating minimization algorithm for criterion (18) different of the standard K-Means. More precisely, one step, updating of the clusters, is performed exactly as in K-Means: just collecting the entities around centers,  $c'$  and  $c''$ , to give  $S_1$  and  $S_2$ , respectively. The other, updating of the center, is somewhat more subtle and based on an equivalent form of (18):

$$(nn_1n_2)^{1/2}L_1 = \sum_{i \in S_1} n_2d(n_1y_i, \bar{c}) + \sum_{i \in S_2} n_1d(-n_2y_i, \bar{c}) \quad (19)$$

where  $\bar{c} = (n_1n_2/n)^{1/2}c$ . The center updated has  $c_v$  equal to  $(n/n_1n_2)^{1/2}\bar{c}_v$  where  $\bar{c}_v$  is the weighted median in the set of reals consisting of  $n_1y_{iv}$  (for  $i \in S_1$ , weight  $n_2$ ) and  $-n_2y_{iv}$  (for  $i \in S_2$ , weight  $n_1$ ), for every  $v \in V$ . The weighted median for a set of decreasing  $x_1, \dots, x_N$  having  $p_1, \dots, p_N$  as their respective weights is defined as the value  $x_a$  ( $a = 1, \dots, N$ ) for which  $\sum_{i < a} p_i = \sum_{i > a} p_i$ ; or, when this equality cannot be achieved, it is an intermediate between those  $x_a$  and  $x_{a+1}$  for which the best approximation of the weight equality is reached.

Regretfully, the  $L_1$  criterion is irrelevant with regard to nominal data.

**Statement 6** *The value of  $L_1$  criterion (18) does not depend on the module-scatter standardized nominal variables.*

**Proof:** Indeed, in a typical case when  $p_v \neq 0.5$ , the weighted median of the column  $y_v$  is zero, which makes the corresponding components in  $c$ ,  $c'$ ,  $c''$  in (18) be zero thus giving the same contribution in either of terms of the criterion.  $\square$

## Acknowledgements

The research was supported by the Office of Naval Research under grants number N00014-93-1-0222 and N00014-96-1-0208 to Rutgers University.

The author thanks Willem Heiser for his support of this work and the referee for comments.

## References

- [1] A. Chaturvedi and J.D. Carroll (1994). An alternating optimization approach to fitting INDCLUS and generalized INDCLUS models. *Journal of Classification* **11** 155-170.
- [2] A.W.F. Edwards and L.L. Cavalli-Sforza (1965). A method for cluster analysis. *Biometrics* **21** 362-375.
- [3] A.K. Jain and R.C. Dubes (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.
- [4] J.B. MacQueen (1967). Some methods for classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium*, Vol. 2, pp. 281-297.
- [5] B.G. Mirkin (1987). Method of principal cluster analysis. *Automation and Remote Control* **48** 1379-1388.
- [6] B. Mirkin (1990). A sequential fitting procedure for linear data analysis models. *Journal of Classification* **7** 167-195.
- [7] B. Mirkin (1996). *Mathematical Classification and Clustering*. Boston-Dordrecht: Kluwer Academic Press.
- [8] B. Mirkin (1997). Concept learning and feature selecting based on square-error clustering. *Machine Learning*. To appear.
- [9] P.H.A. Sneath and R.R. Sokal (1973). *Numerical Taxonomy*. San Francisco: W.H. Freeman.
- [10] S. Van Buuren and W.J. Heiser (1989). Clustering  $N$  objects into  $K$  groups under optimal scaling of variables. *Psychometrika* **54** 699-706.

## EPILOGUE

*Alas, that Spring should vanish with the Rose  
That Youth's sweet-scented Manuscript should close  
The Nightingale that in the Branches sang,  
Ah, whence, and whither flown again, who knows.*

Khayyam Naishapuri: *Persian Astronomer and Poet (1048-1131)*



# Robustness: Where are we now? <sup>1</sup>

Peter J. Huber

*University of Bayreuth, Germany*

*Abstract:* The development of selected robustness concepts since their inception in the 1960's is sketched and their current status is reviewed.

*Key words:* Linear models, estimation, design, regression.

AMS subject classification: 62F35, 62F12, 62A99.

*He looked into the water and saw that it was made up of a thousand thousand thousand and one different currents, each one a different colour, weaving in and out of one another like a liquid tapestry of breathtaking complexity; and Iff explained that these were the Streams of Story, that each coloured strand represented and contained a single tale.*

(Salman Rushdie, *Haroun and the Sea of Stories*, 1990)

## 1 The first ten years

A good case can be made that modern robustness begins in 1960, with the papers by J.W. Tukey on sampling from contaminated distributions, and by F.J. Anscombe on the rejection of outliers. Tukey's paper drew attention to the dramatic effects of seemingly negligible deviations from the model, and it made effective use of asymptotics in combination with the gross error model. Anscombe introduced a seminal insurance idea: sacrifice some performance at the model in order to insure against ill effects caused by deviations from it. Most of the basic ideas, concepts and methods of robustness were invented in quick succession during the following years and

---

<sup>1</sup>First published in *Student* (1995), Vol.1, No.2, 75-86. Reproduced by permission of the Presses Académiques Neuchâtel, Switzerland.

were in place by about 1970.

In 1964 I recast Tukey's general setup into an asymptotic minimax framework and was able to solve it. Important points were: the insistence on finite, but small deviations, the formal recognition that a large error occurring with low probability is a small deviation, and the switch from the then prevalent criterion of relative efficiency to absolute performance. At the same time, I introduced the notion of maximum likelihood type or M-estimators.

Hampel (1968) added the formal definition of qualitative robustness (continuity under a suitable weak topology), infinitesimal robustness in form of the influence function (von Mises derivative of a statistical functional) and the notion of breakdown point.

A nagging early worry had been the possible irrelevancy of asymptotic approaches: conceptually, 1% gross errors in samples of 1000 are entirely different from the same error rate in samples of 5. This worry was laid to rest by Huber (1965, 1968): both for tests and for estimates, there are qualitatively identical and even quantitatively similar exact finite sample minimax robustness results.

The end of the decade saw the first steps of an extension of asymptotic robustness theory beyond location to more general parametric models, namely the introduction of shrinking neighborhoods by C. Huber-Carol (1970), as well as the first attempt at studentizing (Huber, 1970).

The basic methodology for Monte Carlo studies of robust estimators was established in Princeton 1970/71, see Andrews et al. (1972). That study more or less finished off the problem of robustly estimating a single location parameter in samples of size 10 or larger, opening the way for more general multi-parameter problems.

In this paper I shall pick a few of the more interesting robustness ideas and follow the strands of their stories to the present. What has happened to the early strands since their invention? What important new strands have begun in the 1970's and 1980's? I shall avoid technicalities and instead give a reference to a recent survey or monograph, where feasible. Completeness is not intended; for a recent "far from complete" survey of research directions in robust statistics, with more than 500 references, see Stahel's article in Stahel and Weisberg (1991, Part II, p. 243).

## 2 Influence functions and pseudo-values

Among the basic robustness concepts, influence functions have become a standard tool, especially after the comprehensive treatment by Hampel et al. (1986). Also the trick to robustize classical procedures through the

use of pseudo-values is becoming common knowledge, even though it has received only scant coverage in the literature. This trick is that one calculates robust fitted values  $\hat{y}_i$  by iteratively applying the classical procedure to the pseudo-values  $y_i^* = \hat{y}_i + r_i^*$  instead of  $y_i$ . Here, the pseudo-residual  $r_i^* = \psi(r_i)$  is obtained by cutting down the current residual  $r_i = y_i - \hat{y}_i$  with the help of a function  $\psi$  proportional to the desired influence function (i.e. with the  $\psi$ -function defining an M-estimate). For examples see in particular Bickel (1976, p. 167), Huber (1979), and Kleiner, Martin and Thomson (1979). If  $\psi$  is chosen equal rather than merely proportional to the influence function, the classical formulas, when applied to the pseudo-residuals  $r_i^*$  instead of the residuals, yield asymptotically correct error estimates for ANOVA and other purposes (Huber 1981, p. 197).

There have been some very interesting extensions of influence function ideas to time series (Künsch, 1984).

### 3 Breakdown and outlier detection

For a long time, the breakdown point had been a step-child of the robustness literature. The paper by Donoho and Huber (1983) was specifically written to give it more visibility. Recently, I have begun to wonder whether it has given it too much, the suddenly fashionable emphasis on high breakdown point procedures has become counter-productive. One of the most striking examples of the usefulness of the concept can be found in Hampel (1985): the combined performance of outlier rejection followed by the sample mean as an estimate of location is essentially determined by the breakdown of the outlier detection procedure.

### 4 Studentizing

Whenever we have an estimate, we ought to provide an indication of its statistical accuracy, say by giving a 95% confidence interval. This is not particularly difficult if the number of observations is very large, so that the estimate is asymptotically normal with an accurately estimable standard error, or also in one-parameter problems without nuisance parameter, where the finite sample theory of Huber (1968) can be applied.

Otherwise, we end up with a tricky problem of studentization. To my knowledge, there has not been much progress beyond the admittedly unsatisfactory initial paper of Huber (1970). There are not only many open questions with regard to this crucially important problem, it is even open what questions one should ask! A sketch of the principal issues follows.

In the classical normal case, it follows from sufficiency of  $(\bar{x}, s)$  and an



invariance argument that such a confidence interval must take the form

$$(\bar{x} - k_n s / \sqrt{n}, \bar{x} + k_n s / \sqrt{n})$$

with  $k_n$  depending on the sample size, but not on the sample itself. In a well-behaved robust version, a confidence interval might take the analogous form

$$(T - K_n S / \sqrt{n}, T + K_n S / \sqrt{n})$$

where  $T$  is an asymptotically normal robust location estimate and  $S$  is the location invariant, scale equivariant, Fisher consistent functional estimating the asymptotic standard deviation of  $\sqrt{n}T$ , applied to the empirical distribution. In the case of M-estimates for example, we might use

$$S(F)^2 = \frac{S_0^2 \int \psi^2 dF}{(\int \psi' dF)^2}$$

where the argument of  $\psi$ ,  $\psi'$  is  $y = (x - T)/S_0$ , i.e. a robustly centered and scaled version of the observations, say with  $S_0 = MAD$ . If we are interested in 95% confidence intervals,  $K_n$  must approach  $\Phi^{-1}(0.975) \approx 1.96$  for large  $n$ . But  $K_n$  might depend on the sample configuration in a non-trivial, translation and scale invariant way: since we do not have a sufficient statistic, we might want to condition on the configuration of the sample in an as yet undetermined way.

While the distribution of  $\sqrt{n}T$  typically approaches the normal, it will do so much faster in the central region than in the tails, and the extreme tails will depend rather uncontrollably on details of the unknown distribution of the observations. The distribution of  $S$  suffers from similar problems, but here it is the low end which matters. The question is: what confidence levels make sense and are reasonably stable for what sample sizes? For example, given a particular level of contamination and a particular estimate, is  $n = 10$  good enough to derive accurate and robust 99% confidence intervals, or do we have to be content with 95% or 90%? I anticipate that such questions can (and will) be settled with the help of small sample asymptotics, assisted perhaps by configural polysampling (below).

## 5 Shrinking neighborhoods

Direct theoretical attacks on finite neighborhoods work only if the problem is location or scale invariant. But for large samples, most point estimation problems begin to resemble location problems, so it is possible to derive quite general asymptotically robust tests and estimates by letting those neighborhoods shrink at a rate  $n^{-1/2}$  with increasing sample size. This idea

was first exploited by C. Huber-Carol (1970), followed by Rieder, Beran, Millar and others. The final word on this approach is contained in Rieder's book (1994).

The principal motivation clearly is technical: shrinking leads to a manageable asymptotic theory. But there is also a philosophical justification: since the standard goodness-of-fit tests are just able to detect deviations of the order  $O(n^{-1/2})$ , it makes sense to put the border zone between small and large deviations at  $O(n^{-1/2})$ . Larger deviations should be taken care of by diagnostics and modelling, smaller ones are difficult to detect and should be covered (in the insurance sense!) by robustness.

This does not mean that our data samples are supposed to get cleaner if they get larger. But the shrinkage of neighborhood faces us with a dilemma, namely a choice between the alternatives:

- improve the model; or
- improve the data; or
- stop sampling.

Note that adaptive estimation is *not* among the viable alternatives. The problem is not one of reducing statistical variability, but one of avoiding bias, and the ancient Emperor-of-China paradox applies (you can get a fantastically accurate estimate of the height of the emperor by averaging the guesses of 600 million Chinese, most of whom never saw the emperor!).

The asymptotic theory of shrinking neighborhoods is, in essence, a theory of infinitesimal robustness and suffers from the same conceptual drawback as approaches based on the influence function: infinitesimal robustness (bounded influence) does not automatically imply robustness. The crucial point is that in any practical application we have a fixed, finite sample size, and we need to know whether we are inside the range of  $n$  and  $\varepsilon$  for which asymptotic theory yields a decent approximation. This range may be difficult to determine, but the breakdown point often is computable and may be a useful indicator.

## 6 Design

Robustness casts a shadow on the theory of optimal designs: they lose their theoretical optimality very quickly under minor violations of linearity (Huber, 1975) or independence assumptions (Bickel and Herzberg, 1979). I am not aware of much current activity in this area, but the lesson is clear: "Naive" designs usually are more robust and better than "optimal" designs.

## 7 Regression

Back in 1975, the discussants of Bickel (1976) raised interesting criticisms, in particular there were complaints about the multiplicity of robust procedures, and about their computational and conceptual complexity. Bickel fended them off skillfully and convincingly.

There may have been reasons for concern then, but the situation has become worse. Most of the action in the 1980's has been on the regression front. Here is an incomplete list of robust regression estimators: L1 (going back at least to Laplace), M (Huber, 1973), GM (Mallows, 1975), with variants by Hampel, Krasker and Welsch, RM (Siegel, 1982), LMS and LTS (Rousseeuw, 1984), S (Rousseeuw and Yohai, 1984), MM (Yohai, 1987),  $\tau$  (Yohai and Zamar, 1988), SRC (Simpson, Ruppert and Carroll, 1991), and no end is in sight. For an up-to-date review see Davies (1993).

Bickel would not have an easy job now, much of the "Nordic" criticism, unsubstantiated in 1975, seems to be justified now. In any engineering product, an overly rapid sequence of updates sometimes is a sign of vigorous progress, but it can also be a sign of shoddy workmanship, and often it is both. In any case, it confuses the customers and hence is counter-productive...

Robustness has been defined as insensitivity to small deviations from an idealized model. What is this model in the regression case? The classical model goes back to Gauss and assumes that the carrier  $X$  (the matrix  $X$  of the "independent" variables) is error-free.  $X$  may be systematic (as in designed experiments), or haphazard (as in most undesigned experiments), but its rows only rarely can be modelled as being a random sample from a specified multivariate model distribution. Statisticians tend to forget that the elements of  $X$  often are not observed quantities, but are derived from some model (cf. the classical non-linear problems of astronomy and geodesy giving rise to the method of least squares in the first place). In essence, each individual  $X$  corresponds to a somewhat different situation and might have to be dealt with differently. Thus, multiplicity of procedures may lie in the nature of robust regression. Curiously, most of the action seems to have been focused through tunnel vision on just one aspect: safeguard at any cost against problems caused by gross errors in a random carrier.

Over the years, I too had to defend the minimax approach to distributional robustness on many occasions. The salient points of my defense were that the least favorable situation one is safeguarding against, far from being unrealistically pessimistic, is more similar to actually observed error distributions than the normal model, that the performance loss at a true normal model is relatively small, that on the other hand the classically optimal

procedures may lose sorely if the normal model is just slightly violated, and that the hardest problems are not with extreme outliers (which are easy to detect and eliminate), but with what happens on the shoulders of the distributions. Moreover, the computation of robust M-estimates is easy and fast (the last paragraph of this section). Not a single one of these defense lines can be used with the modern “high breakdown point” regression estimates.

A typical cause for breakdown in regression are gross outliers in  $X$ ; while individual such outliers are trivially easy to spot (with the help of the diagonal of the hat matrix), efficient identification of collaborative leverage groups is an open, perhaps unsolvable, diagnostic problem. However, I would advise against treating leverage groups blindly through robustness, they may hide serious design or modeling problems, and there are similar problems even with single leverage points.

The story behind an outlier among the  $X$  (“leverage point”) might for example be:

- a misplaced decimal point,
- an accurate but useless observation, outside of the range of validity of the model.

If the value at this leverage point disagrees with the evidence extrapolated from the other observations, this may be because:

- the outlying observation is affected by a gross error (in  $X$  or in  $y$ ),
- the other observations are affected by small systematic errors (this is more often the case than one might think),
- the model is inaccurate, so the extrapolation fails.

The existence of several, phenomenologically indistinguishable but conceptually different situations with different consequences calls for a diagnostic approach (identification of leverage points or groups), followed by alternative “what if” analyses. This contrasts sharply with simple location estimation, where the observations are exchangeable and a minimax approach is quite adequate (although one may want to follow it up with an investigation of the causes of grosser errors).

At the root of the current confusion is that hardly anybody bothers about stating all of the issues clearly: not only the estimator and a procedure for computing it must be specified, but also the situations for which it is supposed to be appropriate or inappropriate, and criteria for judging estimators and procedures. There has been a tendency to rush into print with rash claims and procedures. In particular, what is meant by the word breakdown? For many of the newer estimates there are unequal-

ified claims that their breakdown point approaches 0.5 in large samples. But such claims tacitly exclude designed situations: if the observations are equally partitioned among the corners of a simplex in  $d$ -space, no estimate whatsoever can achieve a breakdown point above  $1/(2d + 2)$ .

It is one thing to design a theoretical algorithm whose purpose is to prove, for example, that a breakdown point  $1/2$  can be attained in principle, and quite another thing to design a practical version that can be used not merely on small, but also on medium sized regression problems, with a 2000-by-50 matrix or so. This last requirement would seem to exclude all of the recently proposed robust regression estimates.

Some comments on the much maligned “plain vanilla” regression M-estimates are in order. The M-estimate approach is not a panacea (is there such a thing in statistics?), but it is easy to understand, practicable, and considerably safer than classical least squares. While I had had regression problems in mind from the very beginning (cf. Huber 1964, p.74), I had not dared to go into print until 1973, when I believed to understand the asymptotic behavior of M-estimates of regression. A necessary regularity condition for consistency and asymptotic normality of the fitted values is that the maximum diagonal element  $h$  of the hat matrix  $H = X(X^T X)^{-1} X^T$  tends to 0. While watching out for large values of  $h$  does not enforce a high breakdown point, it at least may prevent a typical cause of breakdown. Moreover, with M-estimates of regression, the computing effort for large matrices typically is less than twice what is needed for calculating the ordinary least squares solution. Both calculations are dominated by the computation of a QR or SVD decomposition of the  $X$  matrix, which takes  $O(np^2)$  operations for an  $(n, p)$ -matrix. Since the result of that decomposition can be re-used, the iterative computation of the M-estimate, using pseudo-values, takes  $O(np)$  per iteration with fewer than 10 iterations on average.

## 8 Multivariate problems

Classically, “regression” problems with errors in the independent variables are solved by fitting a least squares hyperplane, that is, by solving a principal component problem and picking the plane belonging to the smallest eigenvalue. It can be argued by analogy that regression problems with potential gross errors in the carrier should be attacked through some version of robust principal components. Thus, multivariate problems, in particular of the principal component type, deserve added attention.

In 1976, Maronna showed that all M-estimates of location and scatter in  $d$  dimensions have a breakdown point  $\varepsilon^* \leq 1/(d + 1)$ . In higher dimensions

this is shockingly low, but then Stahel (1981) and Donoho (1982) independently found estimates based on projection pursuit ideas with a breakdown point approaching  $1/2$  in large samples. The bad news is that with all currently known algorithms the effort for computing those estimates increases exponentially with  $d$ . We might say they break down by failing to give a timely answer!

## 9 Some persistent misunderstandings

Robust methods have persistently been misclassified and pooled with non-parametric and distribution-free methods. They are part of traditional parametric statistics, the only difference is that they do not assume that the parametric model is exactly true.

In part the robustness community itself is to blame for this misunderstanding. In the mid-1970's adaptive estimates - attempting to achieve asymptotic efficiency at all well-behaved error distributions - were thought by many to be the ultimate robust estimates. Then Klaassen (1980) proved a disturbing result on the lack of stability of adaptive estimates. My current conjecture is that an estimate cannot be simultaneously adaptive in a neighborhood of the model and qualitatively robust at the model.

Also the currently fashionable (over-)emphasis of high breakdown points transmits a wrong signal. A high breakdown point is nice to have, if it comes for free, but the potential presence of high contamination usually indicates a mixture model and calls for diagnostics. A thoughtless application of robust procedures might only hide the underlying problem.

There seems to be some confusion between the respective roles of diagnostics and robustness. The purpose of robustness is to safeguard against deviations from the assumptions, in particular against those that are near or below the limits of detectability. The purpose of diagnostics is to find and identify deviations from the assumptions.

The term "robust" was coined by Bayesians (Box and Andersen, 1955). It is puzzling that Bayesian statistics never managed to assimilate the modern robustness concept, but remained stuck with antiquated parametric supermodels. *Ad hoc* supermodels and priors chosen by intuition (personal beliefs) or convenience (conjugate priors) do not guarantee robustness, for that one needs some theory. Compare already Hampel (1973).

## 10 Future directions: small sample problems?

At present, the most interesting and at the same time most promising new methods and open problems have to do with small sample situations.

The Princeton robustness study (Andrews et al., 1972) remains one of the best designed comparative Monte Carlo studies in statistics. Among other things we learned from it how difficult it is to compare the behavior of estimates across entire families of distributions, since small systematic differences of performance easily are swamped by larger random sampling errors. A very sophisticated sampling method thereafter proposed by Tukey is based on the remark that a given sample can occur under any strictly positive density, but it will do so with different probabilities. Thus it must be possible to compare those performances very efficiently by re-using the same sample configurations with different weights when forming Monte Carlo averages. A comprehensive account of this approach is given by Morgenthaler and Tukey (1991).

On the other hand, it was noted by Hampel that a variant of the saddle point method can give fantastically accurate asymptotic approximations down to very small sample sizes (occasionally down to  $n = 1!$ ).

I hope that these approaches in the near future will permit to close several open problems in the area of small samples: studentization, confidence intervals, testing. Embarrassingly, the robustification of the statistics of two-way tables still is wide open. Typically there are so few degrees of freedom per cell that ordinary asymptotic approaches are out of the question, maybe some version of small sample asymptotics may help here too.

## References

- [1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton: Princeton Univ. Press.
- [2] Anscombe, F.J. (1960). Rejection of outliers. *Technometrics* **2**, 123-147.
- [3] Bickel, P.J. (1976). Another look at robustness: A review of reviews and some new developments. *Scand. J. Statist.* **3**, 145-168.
- [4] Bickel, P.J., and Herzberg, A.M. (1979). Robustness of design against autocorrelation in time. *Ann. Statist.* **7**, 77-95.
- [5] Box, G.E.P., and Andersen, S.L. (1955). Permutation theory in the derivation of robust criteria and the study of departure from assumption. *J. R. Statist. Soc. B* **17**, 1-34.
- [6] Davies, P.L. (1993). Aspects of robust linear regression. *Ann. Statist.* **21**, 1843-1899.
- [7] Donoho, D.L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Harvard University, Cambridge, MA.
- [8] Donoho, D.L., and Huber, P.J. (1973). The notion of breakdown point.

In *A Festschrift for Erich L. Lehmann*, Eds. P. J. Bickel, K.A. Doksum and J.L. Hodges. Wadsworth, Belmont, CA.

- [9] Hampel, F.R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis, Univ. of Calif., Berkeley.
- [10] Hampel, F.R. (1973). Robust estimation: A condensed partial survey. *Z. Wahr. verw. Geb.* **27**, 87-104.
- [11] Hampel, F.R. (1985). The breakdown point of the mean combined with some rejection rules. *Technometrics* **27**, 95-107.
- [12] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics. The Approach Based on Influence*. New York: Wiley.
- [13] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.
- [14] Huber, P.J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* **36**, 1753-1758.
- [15] Huber, P.J. (1968). Robust confidence limits. *Z. Wahr. verw. Geb.* **10**, 269-278.
- [16] Huber, P.J. (1970). Studentizing robust estimates. In *Nonparametric Techniques in Statistical Inference*, Ed. M. L. Puri. Cambridge: Cambridge University Press.
- [17] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799-821.
- [18] Huber, P.J. (1975). Robustness and designs. In *A Survey of Statistical Design and Linear Models*, Ed. J. N. Srivastava. Amsterdam: North Holland.
- [19] Huber, P.J. (1979). Robust smoothing. In *Proc. ARO Workshop on Robustness in Statistics*, Eds. R. N. Launer and G. N. Wilkinson. New York: Academic Press.
- [20] Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.
- [21] Huber-Carol, C. (1970). Etude asymptotique de tests robustes. Ph.D. thesis, ETH, Zürich.
- [22] Klaassen, C. (1980). Statistical performance of location estimators. Ph.D. thesis, Mathematisch Centrum, Amsterdam.
- [23] Kleiner, B., Martin, R.D., and Thomson, D.J. (1979). Robust estimation of power spectra. *J. R. Statist. Soc. B* **41**, 313-351.
- [24] Künsch, H. (1984). Infinitesimal robustness for autoregressive processes. *Ann. Statist.* **12**, 843-863.
- [25] Mallows, C.L. (1975). On some topics in robustness. Tech. Memorandum, Bell Telephone Laboratories, Murray Hill, NJ.
- [26] Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4**, 51-67.



- [27] Morgenthaler, S. and Tukey, J.W. (1991). *Configural Polysampling*. New York: Wiley.
- [28] Rieder, H. (1994). *Robust Asymptotic Statistics*. Berlin: Springer.
- [29] Rousseeuw, P.J. (1984). Least median of squares regression. *J. Am. Statist. Assoc.* **79**, 871-880.
- [30] Rousseeuw, P.J., and Yohai, V.J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*, Eds. J. Franke, W. Härdle and R.D. Martin. Lecture Notes in Statistics 26. New York: Springer.
- [31] Siegel, A.F. (1982). Robust regression using repeated medians. *Biometrika* **69**, 242-244.
- [32] Simpson, D.G., Ruppert, D., and Carroll, R.J. (1992). On one-step GM-estimates and stability of inferences in linear regression. *J. Am. Statist. Assoc.* **87**, 439-450.
- [33] Stahel, W.A. (1981). Breakdown of covariance estimators. Research Report 31, Fachgruppe für Statistik, ETH Zürich.
- [34] Stahel, W., and Weisberg, S. (Eds.) (1991). *Directions in Robust Statistics and Diagnostics* Part I and II. The IMA Volumes in Mathematics and its Applications, Vols. 33-34. New York: Springer.
- [35] Student (1927). Errors of routine analysis. *Biometrika* **19**, 151-164.
- [36] Tukey, J.W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*, Ed. I. Olkin. Stanford: Stanford Univ. Press.
- [37] Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15**, 642-656.
- [38] Yohai, V.J., and Zamar, R.H. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *J. Am. Statist. Assoc.* **83**, 406-413.

# Abstracts

## Asymptotic properties of $L_1$ estimators in a multi-stage dose-response model: A Monte Carlo study

Carmen D.S. André, *University of São Paulo, Brazil*

Subhash C. Narula, *Virginia Commonwealth University, Richmond, USA*

*Abstract:* The single-stage and two-stage dose response models are frequently used in practical applications. The maximum likelihood and the least squares principles are often used to estimate the unknown parameters of the model. It has been shown that these methods are sensitive to outliers in the data. The minimum sum of absolute errors MSAE (or  $L_1$ ) criterion is more resistant to outliers than these popular procedures. However, at present not much is known about the statistical properties of the MSAE estimators of the parameters of the multistage dose-response model. In this paper, our objective is to study asymptotic properties and distribution of the MSAE estimators of the single-stage and two-stage dose-response models by simulation and to find the smallest sample size for which we may use the asymptotic distribution to draw statistical inferences about the parameters. We also give an approximate expression for the variance of these estimators when their asymptotic distribution follows a multinormal distribution.

## The $L^1$ -norm and interlaboratory tests

P. L. Davies, *University of Essen, Germany*

*Abstract:* The form of interlaboratory test we consider is that where each of  $I$  laboratories returns exactly one reading for each of  $J$  samples. Such a test may be described by the random effects model

$$X_{ij} = X_i + a_j + \epsilon_{ij}, \quad 1 \leq i \leq I, \quad 1 \leq j \leq J.$$

The  $X_i$  represent the laboratory effects, the  $a$  the sample contaminations and the  $\epsilon_{ij}$  the measurement errors. The problem is to identify outlying observations and outlying laboratories. As we have only one observation per cell it is commonly believed that it is not possible to detect outliers or, equivalently, non-additivity. As shown in Terbeck and Davies (1996) this is not correct and so called unconditionally identifiable outlier patterns may be found by the  $L^1$ - or an appropriate  $M$ -functional. The results

of Terbeck and Davies are improved in certain respects and then applied to the random effects model. The method is applied to a real data set considered by Lischer (1993) which is concerned with the determination of lead in sewage sludge.

### **Multivariate $L_1$ mean**

**Yadolah Dodge and Valentin Rousson,**  
*University of Neuchâtel, Switzerland*

*Abstract:* The center of a univariate data set  $\{x_1, \dots, x_n\}$  can be defined as the point  $\mu$  that minimizes the norm of the vector of distances  $\mathbf{y}' = (|x_1 - \mu|, \dots, |x_n - \mu|)$ . As the median and the mean are the minimizers of respectively the  $L_1$ - and the  $L_2$ -norm of  $\mathbf{y}$ , they are two alternatives to describe the center of a univariate data set. The center  $\mu$  of a multivariate data set  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  can also be defined as minimizer of the norm of a vector of distances. In multivariate situations however, there are several kinds of distances. In this paper, we consider the vector of  $L_1$ -distances  $\mathbf{y}'_1 = (\|\mathbf{x}_1 - \mu\|_1, \dots, \|\mathbf{x}_n - \mu\|_1)$  and the vector of  $L_2$ -distances  $\mathbf{y}'_2 = (\|\mathbf{x}_1 - \mu\|_2, \dots, \|\mathbf{x}_n - \mu\|_2)$ . We define the  $L_1$ -median and the  $L_1$ -mean as the minimizers of respectively the  $L_1$ - and the  $L_2$ -norm of  $\mathbf{y}_1$ ; and then the  $L_2$ -median and the  $L_2$ -mean as the minimizers of respectively the  $L_1$ - and the  $L_2$ -norm of  $\mathbf{y}_2$ . In doing so, we obtain four alternatives to describe the center of a multivariate data set. While three of them have been already investigated in the statistical literature, the  $L_1$ -mean appears to be a new concept. Contrary to the  $L_1$ -median, the  $L_1$ -mean is proved to be unique in almost all situations. In order to compare these multivariate medians and means, we use the rule of the net advantage coefficient introduced by Stavig and Gibbons (1977). A simulation study shows that the  $L_1$ -mean performs well, especially for data sets drawn from bivariate Laplace distribution.

### **The information for the direction of dependence in $L_1$ regression**

**Yadolah Dodge,** *University of Neuchâtel, Switzerland*  
**Joe Whittaker,** *Lancaster University, U.K.*

*Abstract:* An  $L_1$  regression model for a response variable  $X_2$  is to suppose that the conditional distribution of  $X_2$  given  $X_1$  is Laplace, and that the marginal distribution of the explanatory variable  $X_1$  is also Laplace.

We show that there is information to distinguish the direction of dependence  $X_1$  and  $X_2$ ; or equivalently to distinguish between the models in which  $X_1$  is dependent on  $X_2$ , and  $X_2$  is dependent on  $X_1$ . This is not true for  $L_2$  regression based on the Normal distribution.

## **Dimension choice for sliced inverse regression based on ranks**

**Louis Ferré**, *Université Paul Sabatier, Toulouse, France*

*Abstract:* Sliced Inverse Regression is a method for reducing the dimensionality in multivariate non parametric regression problems. While the selection of the dimensionality has been investigated for the original version, no solution has been proposed for Hsing and Carroll (1992) approach based on order statistics and associated concomitant variables. By using model selection approaches, we propose here two ways for selecting the dimensionality by estimating a loss function: first, a direct estimation is proposed and, then a Jack-Knifed estimate is investigated. Finally, the rank version is compared to classical SIR on a real life data set.

## **$L_1$ -norm and $L_2$ -norm methodology in cluster analysis**

**Allan D. Gordon**, *University of St Andrews, North Haugh, Scotland*

*Abstract:* An overview is provided of the use of  $L_1$ -norm and  $L_2$ -norm methodology in cluster analysis. Topics covered include dissimilarity measures, partitions, fuzzy classifications, hierarchical classifications, and consensus classifications.

## **Some issues in the applications of conditional quantile functions**

**Xuming He**, *University of Illinois, USA*

*Abstract:* Conditional quantile functions are useful in a variety of applications. Regression quantiles for linear models have been recently extended to semiparametric and nonparametric models. Further investigations are needed for both the statistical theory and computations. In this paper, I attempt to raise two questions that I believe are important to build a

solid foundation for the applications of quantile regression. They focus on nearly extreme quantiles and the problem of crossing in estimated quantile functions.

## **The median function on structured metric spaces**

**F. R. McMorris**, *University of Louisville, USA*

*Abstract:* When  $(X, d)$  is a finite metric space and  $\pi = (x_1, \dots, x_k) \in X^k$ , a median for  $\pi$  is a element  $x$  of  $X$  for which  $\sum_{i=1}^k d(x, x_i)$  is minimum. The function that returns the set of all medians for any tuple  $\pi$  is called the median function on  $X$ . A brief survey is given of some of the results concerning the median function, starting with an arbitrary metric space and finishing with the case where  $X$  is a set of hypergraphs and  $d$  is the metric based on the  $L_1$ -norm. A simplistic maximum likelihood interpretation for the median function is also given.

## **Least absolute value estimation of a linear functional model**

**Édina Shisue Miazaki**, *Statistika Consultoria, Campinas, Brazil*  
**Gabriela Stangenhau**, *Universidade de Brasília, Brazil*

*Abstract:* This paper presents two robust  $L_1$  based estimators for the parameters of a simple linear functional relationship (SLFR). The maximum likelihood estimation when the errors follow a double exponential distribution and the weighted  $L_1$  estimation are solved as non-linear optimization problems. The least median of squares estimates are proposed as starting values and the scale measures of the errors are based on the MAD. Both methods are resistant to outlying observations and the weighted  $L_1$  estimator is resistant to leverage points. Real examples illustrate the methods.

## **ANOVA - models: A Bayesian analysis**

**Wolfgang Polasek and Shuangzhe Liu**, *University of Basel, Switzerland*

*Abstract:* The 1-way and 2-way ANOVA are formulated as Bayesian linear models with conjugate prior distributions. The classical case is treated



as a special one using matrix generalized (g-) inverses leading to so-called  $OLS^-$  and  $OLS^+$  estimates of the rank deficient ANOVA model. The 2-way ANOVA model without interactions can also be estimated in a 2-step procedure.

## **Robustifying growth curve model estimation**

**Gabriela Stangenhau**, *Statistika Consultoria, Campinas, Brazil*

**Elisete C. Quintaneiro Aubin**, *Universidade de São Paulo, Brazil*

*Abstract:* A robustified version of the parameter matrix estimators in the standard growth curve model obtained via the Potthoff-Roy transformation is presented. The asymptotic distribution of the robust estimators is derived and the estimation of their variance-covariance matrix is discussed.

## **Fitting $L_2$ norm classification models to complex data sets**

**Maurizio Vichi**, *University "G.D'Annunzio" di Chieti, Pescara, Italy*

*Abstract:* In this paper methodologies for fitting classification models (dendrograms and partitions) to two and three-way arrays of dissimilarities minimizing a  $L_2$  norm loss function are examined. A new algorithm for fitting several hierarchical classifications to quite large three-way arrays is also discussed.

## **Applications of mathematical programming in $L_1$ -estimation of nonlinear models**

**Jinde Wang**, *Nanjing University, China*

*Abstract:* In this paper we review the results in  $L_1$ -estimation of nonlinear models obtained by applying mathematical programming techniques. We describe briefly the ways to find the asymptotic distribution, the approximate representation and to treat dependent random error cases and inequality-constrained cases. With these results one can conclude that mathematical programming is a suitable tool for studying  $L_1$ -estimation problems.

## Some contributions to M-estimation in regression models

**Lincheng Zhao,**

*University of Science and Technology of China, Hefei, China*

*Abstract:* In this paper, we briefly survey some contributions to asymptotic theory on M-estimation in a linear model and least absolute deviations (LAD) estimation in a censored regression model (known as the Tobit model), as well as on the relevant test criteria in ANOVA in the above models. As a general approach on statistical data analysis, asymptotic theory of M-estimation in regression models has received extensive attention. In recent years, the author and some of his cooperators worked on this field and obtained some new results. In this paper we briefly introduce some of them and the related work in the literature. As a special case, the minimum  $L_1$ -norm ( $ML_1N$ ) estimation, also known as the least absolute deviations (LAD) estimation, plays an important role and is of special interest. Considering this point, we will pay much attention to them as well. Our topics concern the usual linear model and a censored regression model, known as the Tobit model in econometric research.

# Contributors

Adcock Christopher, School of Economic and Business Studies, Univ. of Westminster, 309 Regent Street, W1R 8AL London, U.K.

Agullo Jose, Dept. Fundamentos del Analisis Economico, Universidad de Alicante, 03080 Alicante, SPAIN

Andre D.S. Carmen, Instituto de Matematica e Estatistica da USP, Rua do Matao 1010, C.P. 66281 - Agencia Cidade de Sao Paulo, CEP 05315-970 Sao Paulo-SP, BRAZIL

Arabie Phipps, Graduate School of Management, Rutgers University, 92, New Street, NJ 07102-1895 Newark, USA

Aubin Elisete, Instituto de Matematica e Estatistica da USP, Rua do Matao, 1010, CEP, 05508-900 Sao Paulo, SP, BRAZIL

Bassett Gilbert W., Dept. of Economics, Univ. Illinois-Chicago, 601 S. Morgan St, Rm 2103, 60607-7121 Chicago/ Illinois, USA

Brown Bruce M., University of Tasmania-Math, P.O. Box 252 C, 7001 Hobart Tasmania, AUSTRALIA

Bura Efstathia, Dept. of Statistics, George Washington Univ., DC 20052 Washington, USA

Cadarso-Suarez Carmen, Dept de Estadistica e I.O., Seccion de Bioestadistica, Facultad de Medicina, Universidad de Santiago, La Coruna Santiago de Compostela, SPAIN

Carroll Douglas, Faculty of Management/ Marketing, Rutgers Univ., 81 New St, NJ 07102-1895 Newark, USA

Chakraborty Biman, Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, 203 B.T. Road, 700035 Calcutta, INDIA

Chaturvedi Anil, Technology Consultant, AT&T Labs, Room 5C-134, 600 Mountain Avenue, NJ 07974 Murray Hill, USA

Chaudhuri Probal, Theoretical Stats & Math., Indian Stat. Inst., 8 Ramlal Agarwalla Lane, 700050 Calcutta, INDIA

Chepoi Victor, Laboratoire de Biomathematiques, Faculté de Médecine, Université d'Aix Marseille II, 27 Bd Jean Moulin, 13385 Marseille Cedex 5, FRANCE

Chiaromonte Francesca, International Institute for Applied Systems Anal-



ysis, A-2361 Laxenburg, AUSTRIA

Cogneau Daniel, Faculté de Médecine, Laboratoire de Biomathématique, 27, Bd. Jean Moulin, 13385 Marseille, FRANCE

Davies Laurie, Univ. Gesamthochschule Essen, FB6 Maths, Universitätsstrasse 3, 4300 Essen 1, GERMANY

de Una Alvarez Jacobo, Depto de Estadística e I.O., Campus Lagoas-Marcosende, Universidad de Vigo, 86200 Vigo, SPAIN

Dodge Yadolah, Groupe de Statistique, Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, SWITZERLAND

Farebrother Richard William, 11 Castle Road, Bayston Hill, Shrewsbury, SY3 ONF Shropshire, U.K.

Fernholz Luisa, Statistics Dept., Temple Univ., Speakman Hall, PA 19122 Philadelphia, USA

Ferré Louis, Laboratoire de Statistique, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, FRANCE

Fichet Bernard, Faculté de Médecine, Laboratoire de Biomathématique, 27, Bd. Jean Moulin, 13385 Marseille, FRANCE

Fitzenberger Bernd, Department of Economics, and Statistics, University of Konstanz, P.O. Box 5590 < D 139 >, D-78434 Konstanz, GERMANY

Fraiman Ricardo, Centro de Matematica, Universidad de la Republica-Uruguay, Eduardo Acevedo 1139, Montevideo, URUGUAY

Gordon Allan D., Mathematical Institute, University of St Andrews, North Haugh, St Andrews KY16 9ss, Fife, SCOTLAND

Gonzalez-Manteiga Wenceslao, Departamento de Estadística, Facultad de Matematicas, Universidad de Santiago de Compostela, La Coruna, SPAIN

Gutenbrunner Cornelius, Uni-Klinikum, Abteilung für Kinder und Jugendpsychiatrie, Hans-Sachs Strasse 6, 35033 Marburg, GERMANY

Hall Peter, Dept. of Statistics, A.N.U., G.P.O. Box 4, A.C.T. 2601 Canberra, AUSTRALIA

Hallin Marc, Institut de Statistique, Campus de la Plaine, Univ. Libre de Bruxelles, C.P. 210, B-1050 Bruxelles, BELGIUM

He Xuming, Dept. of Statistics, Univ. of Illinois-Champaign, 725 S Wright St, IL 61820-5714 Champaign, USA

Hettmansperger Tom, Pennsylvania State Univ., Dept. of Statistics, 317 Classroom Building, PA 16802-2111 University Park, USA

Huber, Peter J., LSt. Mathematik, Universität Bayreuth, Universitätsstrasse 30, 95440 Bayreuth, GERMANY

Hubert Lawrence, Dept. of Psychology (and Dept. of Statistics), University of Illinois-Champaign, 603 East Daniel Street, IL 61820-6232 Champaign, USA

Hubert Mia, Dept of Mathematics, Universitaire Inst. Antwerpen, Universiteitsplein 1, 2610 Wilrijk, BELGIUM

Hurst Simon R., Centre for Financial Mathematics, School of Mathematical Sciences, Australian National Univ., ACT 0200 Canberra, AUSTRALIA

Huskova Marie, Charles Univ., Dept. of Probability and Stat., Sokolovska 83, 18600 Prague 8, CZECH REPUBLIC

Jureckova Jana, Charles Univ., Dept. of Probability and Stat., Sokolovska 83, 18600 Prague 8, CZECH REPUBLIC

Kariya Takeaki, Institute of Economic Research, Hitotsubashi Univ. Kunitachi, 186 Tokyo, JAPAN

Klebanov L.B., Institute of Mathematical Geology, St Petersburg, RUSSIA

Knight Keith, Dept. of Statistics, Univ. of Toronto, 100 St-George Street, M5S 1A1 Toronto, CANADA

Koenker Roger, Dept. of Economics, Univ. of Illinois, 1206 S 6th St, IL 61820-6915 Champaign, USA

Liu Regina, Dept. of Statistics-Hill Cntr, Rutgers University, NJ 08855 Piscataway, USA

Liu Shuangzhe, Institute of Statistics and Econometrics, Holbeinstrasse 12, 4051 Basel, SWITZERLAND

McMorris F.R., Univ. of Louisville, Dept. of Mathematics, KY 402 92 Louisville, USA

McKean Joe, Dept. of Math & Statistics, Western Michigan Univ., MI 49 008 Kalamazoo, USA

Meade Nigel, The Management School, Imperial College, 53 Princes Gate, SW7 2PG London, U.K.

Meloche Jean, Dept. of Statistics, 2021 West Mall, Univ. of British Columbia, BC V6T 1W5 Vancouver, CANADA

Meulman Jacqueline J., Department of Data Theory, Leiden University, P.O. Box 9555, 2300 RB Leiden, THE NETHERLANDS

Miazaki Edina Shisue, C.P. 04393, 70919-970 Brasilia, D.F., BRAZIL

Mirkin Boris G., Division of Theoretical Bioinformatics (Abt 0815), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, GERMANY

Mizera Ivan, Department of Probability and Statistics, Comenius University, Mlynska dolina, 84215 Bratislava, SLOVAKIA

Moreno Elias, Department of Statistics, Faculty of Science, University of Granada, 18071 Granada, SPAIN

Morgenthaler Stephan, EPFL, Dept. of Mathematics, 1015 Lausanne, SWITZERLAND

Möttönen Jyrki, Department of Mathematical Sciences, University of Oulu, 90570 Oulu, FINLAND

Müller Christine, Freie Universität Berlin, Fachbereich Math. & Informatik, WE1, Arnimalle 2-6, 14195 Berlin, GERMANY

Narula Subash C., School of Business, Virginia Commonwealth Univ., 1015 Floyd Avenue, VA 23284-4000 Richmond, USA

Nolan Deborah A., Dept. of Statistics, 367 Evans Hall, Univ. of California, CA 94720-3860 Berkeley, USA

Nyquist Hans, Swedish Univ. of Agriculture, Dept. of Biometry, 901 83 Umea, SWEDEN

Oja Hannu, Dept. of Applied Maths and Statistics, Linnanmaa, Univ. Oulu, 90570 Oulu 57, FINLAND

Platen Eckhard, Centre for Financial Mathematics, School of Mathematical Sciences, The Australia National University, 0200 Canberra ACT, AUSTRALIA

Polasek Wolfgang, Institute of Statistics and Econometrics, Holbeinstrasse 12, 4051 Basel, SWITZERLAND

Portnoy Steve, Dept. of Statistics, Univ. of Illinois-Champaign, 725 S Wright St, IL 61820-5714 Champaign, USA

Raimondo Marc, Australian National University, Department of statis-

tics, G.P.O. Box 4, A.C.T. 2601 Canberra, AUSTRALIA

Rousseeuw Peter, Dept. of Mathematics, Universitaire Inst. Antwerpen, Universiteitsplein 1, 2610 Wilrijk, BELGIUM

Rousson Valentin, Groupe de Statistique, Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, SWITZERLAND

Sheather Simon, Australian Graduate School of Management, University of New South Wales, 2033 Kensington NSW, AUSTRALIA

Stangenhuis Gabriela, Rua Americo de Campos 133, 13083-040 Campinas SP, BRAZIL

Stute Winfried, Math. Institut, Univ. Giessen, Arndtastr. 2, 35392 Giessen, GERMANY

Terui Nobuhiko, Tinbergen Institute, Rotterdam, Burg Oudlaan 50, PA 3062 Rotterdam, THE NETHERLANDS

Vichi Maurizio, Universit "G. D'Annunzio" di Chieti, Dept. of Metodi Quantitativi e Teoria Economica, Viale Pindaro 42, 65127 Pescara, ITALY

Wang Jinde, Dept. of Mathematics, Nanjing Univ., 22, Hankou Road, 210008 Nanjing, CHINA

Wetzel Nate, Dept. of Mathematical Sciences, Binghamton Univ., NY 13902-6000 Binghamton, USA

Whittaker Joe, Dept of Mathematics, Lancaster University, LA1 4YF Lancaster, U.K.

Zhao Lincheng, Department of Mathematics, Univ. of Science & Technology, HEFEI, 230026 Anhui, CHINA

Zwanzig Silvelyn, Institut für Mathematische Stochastik, Universität Hamburg, Bundesstrasse 55, 20146 Hamburg, GERMANY

ISBN 0-940600-43-9